Неравенства концентрации в линейных моделях большой размерности

Голубев Г.К.

CNRS, Université de Provence и ИППИ РАН

19.05.2012

Основная задача статистики – восстановить вероятностное распределение случайного объекта по его реализации, которая представляет собой, как правило, некоторый вектор $Y \in \mathbb{R}^n$. Эта задача решается при помощи трех основных объектов:

- *Модель наблюдений (вероятностная априорная информация о наблюдениях)*
  Например, одной из самых широко распространенных моделей в статистике является линейная модель

$$Y = X\mu + \epsilon;$$

  здесь
  - $\mu \in \mathbb{R}^p$ – неизвестный параметр,
  - $Y \in \mathbb{R}^n$ – наблюдения,
  - X – известная $n \times p$-матрица,
  - $\epsilon \in \mathbb{R}^n$ – белый гауссовский шум с известной или неизвестной интенсивностью σ (ϵ_i н.о.р. гауссовские с.в. с $\mathbf{E}\epsilon_i = 0$ $\mathbf{E}\epsilon_i^2 = \sigma^2$).

- *Априорная информация о параметрах модели.*
  Например,
  - $$\|\mu\|^2 = \sum_{k=1}^{p} \mu_k^2 \leq E;$$

    здесь E - параметр, который может быть как известен, так и неизвестен.
  - μ – случайный вектор с независимыми $\mathcal{N}(0, S^2)$ компонентами; здесь дисперсия S^2 может быть опять же как известна, так и неизвестна.
  - $$\|\mu\|_1 = \sum_{k=1}^{p} |\mu_k| \leq E.$$
  - μ – случайный вектор с независимыми компонентами с плотностью

$$p_S(x) = \frac{1}{2S} \exp\left(-\frac{|x|}{S}\right), \ S \in \mathbb{R}^+.$$

- *Мера качества оценивания.*
  В статистике, как правило, качество некоторой оценки $\hat{\mu}(Y)$ измеряется ее риском

$$\mathbf{E}l(\hat{\mu}(Y), \mu),$$

  где

  - $\mathbf{E}$ – усреднение по мере, порожденной наблюдениями Y
  - $l(\cdot, \cdot) : \mathbf{R}^p \times \mathbf{R}^p \to \mathbb{R}^+$ – функция потерь. Функция потерь определяется статистическим смыслом рассматриваемой задачи. Например,

$$l(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2 \text{ или } l(\hat{\mu}, \mu) = \|X\hat{\mu} - X\mu\|^2.$$

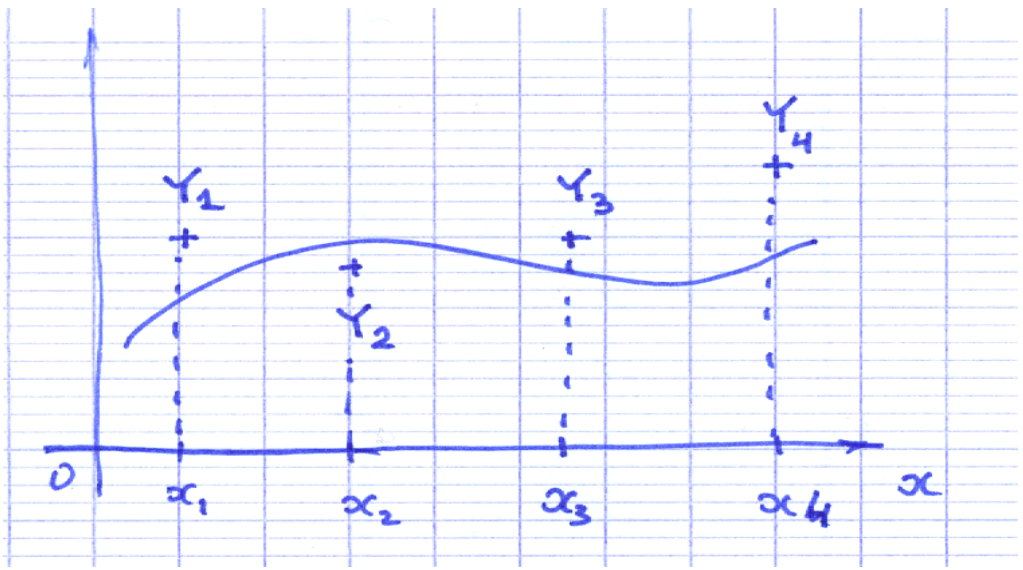
    С вероятностной точки зрения риск – это мера концентрации оценки $\hat{\mu}(Y)$.

Полиномиальная регрессия

Предположим, что мы имеем наблюдения

$$Y_i = f(x_i) + \sigma\xi_i,\ x_i \in [0,1],\ i = 1, \dots, n,$$

где ξ_i — н.о.р. $\mathcal{N}(0,1)$, а $f(x),\ x \in [0,1]$ — неизвестная функция, которую мы хотим восстановить по наблюдениям Y.



Предположим, что $f(x)$ является гладкой функцией и может быть приближена некоторым полиномом

$$f(x) = \sum_{k=0}^{n-1} x^k \mu_k;$$

здесь μ_k — неизвестные величины, которые мы должны оценить по наблюдениям Y_i, $i = 1, \ldots, n$.
Чтобы свести эту задачу к линейной модели, будем считать, что столбцами матрицы X являются векторы

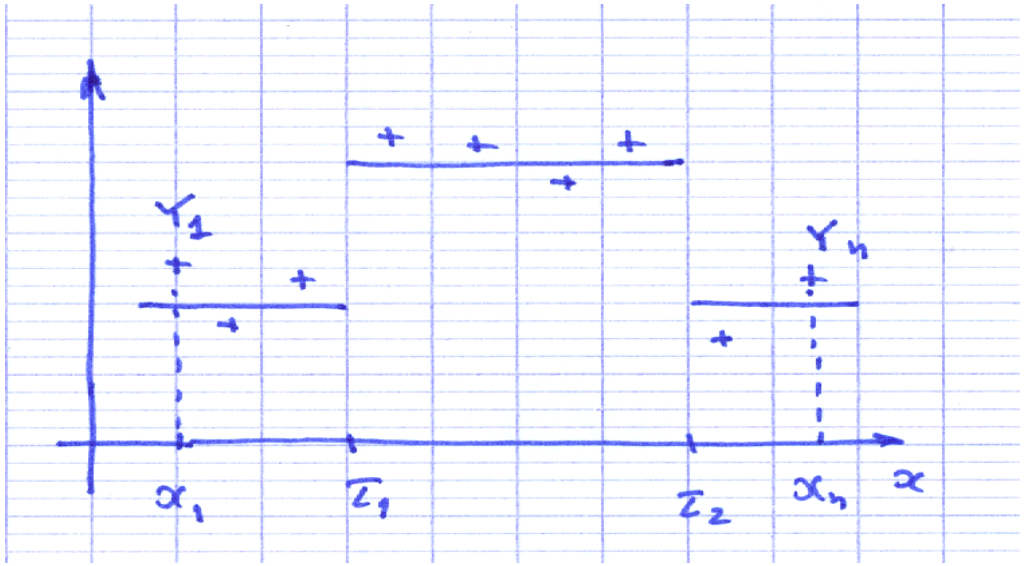
$$P_k = (x_1^k \ldots, x_n^k)^\top \quad \text{т. е.} \quad X = [P_1, \ldots, P_n].$$

Априорная информация о векторе μ:

$$\mu_k = 0 \quad \text{при всех} \quad k > Q,$$

где $Q \in \{1, \ldots, n-1\}$ — некоторое неизвестное целое число (степень полинома).

Сглаживание кусочно-постоянных сигналов



Предположим, что мы наблюдаем

$$Y_i = f_i + \sigma\xi_i, \quad i = 1, \ldots, n,$$

где f_i — кусочно-постоянный вектор, т.е.

$$f_i = c_k \quad \text{при} \quad \tau_k \le i < \tau_{k+1}.$$

Предполагается, что целые числа $\tau_k,\ k = 1, \ldots, M$ упорядочены и таковы, что $\tau_1 = 1$ и $\tau_M = n + 1$.
Будем считать, что ни эти числа, ни $c_k,\ k = 1, \ldots, M - 1$, ни M нам неизвестны. Чтобы перейти к линейной модели возьмем в качестве X нижне-треугольную матрицу:

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Тогда

$$f = X\mu,$$

где μ — разреженный вектор, т.е. вектор у которого большинство компонент равны 0.
Априорная информация о μ, отражающая его разреженность может иметь следующий вид:

$$\|\mu\|_\circ \le E, \quad \text{где} \quad \|\mu\|_\circ = \sum_{i=1}^{n} \mathbf{1}\{|\mu_i| > 0\}$$

или

$$\|\mu\|_1 \le E, \quad \text{где} \quad \|\mu\|_1 = \sum_{i=1}^{n} |\mu_i|.$$

Вопрос о том, какую априорную информацию мы должны использовать, связан с вычислительной сложностью решения оптимизационных задач

$$\hat{\mu}_\circ = \arg\min_\mu \left\{ \|Y - X\mu\| + \sigma S \|\mu\|_\circ \right\},$$

$$\hat{\mu}_1 = \arg\min_\mu \left\{ \|Y - X\mu\| + \sigma S \|\mu\|_1 \right\};$$

Здесь S — параметр регуляризации.
Первая задача решается за $O(\exp(n))$ операций, вторая за $O(n^3)$.
Второй метод в литературе называется LASSO (Least Absolute Shrinkage and Selection Operator).

Интегральные уравнения

Предположим, что мы хотим решить интегральное уравнение

$$h(t) = \int_0^1 A(t,s)x(s)\,ds, \quad t \in [0,1].$$

Выберем две системы ортонормальных функций $\phi_k(\cdot)$, $\psi_k(\cdot)$. Тогда представляя

$$x(s) = \sum_{k=1}^{\infty} \mu_k \phi_k(s), \quad Y_k = \langle h, \psi_k \rangle,$$

приходим к уравнению для μ

$$Y = X\mu + \sigma\xi, \quad X_{ij} = \int_0^1 \int_0^1 \psi_i(t)A(t,s)\phi_j(s)\,ds\,dt.$$

Естественная априорная информация о μ: $\|\mu\|^2 \le E$.

При оценивании параметра μ по наблюдениям Y мы хотим найти оценку $\hat{\mu}^*(Y)$ с минимальным риском т.е. такую, что

$$\mathbf{E}l(\hat{\mu}^*, \mu) \leq \mathbf{E}l(\hat{\mu}, \mu) \text{ для всех } \mu \in \mathbb{R}^p \text{ и любой оценки } \hat{\mu}.$$

В отличии от теории вероятностей, в которой, как правило, рассматривается задача изучения концентрации заданной функции $\hat{\mu}(Y)$, в статистике наибольший интерес представляет обратная задача, а именно, поиск функции $\hat{\mu}(Y)$, обладающей наилучшей концентрацией.

Поскольку риск $\mathbf{E}l(\hat{\mu}, \mu)$ зависит, как правило, от μ, эта идея нереализуема, т.к. мы не можем, вообще говоря, сравнивать функции.
Сравнивать можно числа и поэтому довольно естественно определить наилучшую оценку $\hat{\mu}_\pi$, как оценку, для которой

$$\int_{\mathbb{R}^p} \mathbf{E}l(\hat{\mu}_\pi, \mu)\pi(\mu)\, d\mu \leq \int_{\mathbb{R}^p} \mathbf{E}l(\hat{\mu}, \mu)\pi(\mu)\, d\mu \text{ для любой оценки } \hat{\mu};$$

здесь функция $\pi(\cdot) \geq 0$, называемая априорной плотностью распределения параметра μ такова, что

$$\int_{\mathbb{R}^p} \pi(\mu)\, d\mu = 1.$$

При таком этом подходе к сравнению оценок мы считаем, что μ — случайная величина с плотностью $\pi(\cdot)$.

Баейсовская оценка $\hat{\mu}_\pi(Y)$ вычисляется следующим образом:

$$\hat{\mu}_\pi(Y) = \arg\min_{\mu'} \int_{\mathbb{R}^d} l(\mu', \mu) p_\mu(Y) \pi(\mu)\, d\mu;$$

здесь $p_\mu(\cdot)$ – плотность распределения наблюдений $p_\mu(\cdot)$.

Если рапределение (Y, μ) гауссовское и $l(\cdot, \cdot)$ квадратична, то $\hat{\mu}_\pi(Y)$ — линейна по Y.

За качество оценки $\hat{\mu}_\pi$ отвечают два объекта

1. плотность распределения наблюдений $p_\mu(Y)$,
2. априорная плотность распределения неизвестного параметра $\pi(\cdot)$.

С этими плотностями связаны две информационные матрицы Фишера

$$I_Y = \mathbf{E}\left(\frac{\partial}{\partial\mu}\log[p_\mu(Y)]\right)\left(\frac{\partial}{\partial\mu}\log[p_\mu(Y)]\right)^\top,$$
$$I_\pi = \mathbf{E}\left(\frac{\partial}{\partial\mu}\log[\pi(\mu)]\right)\left(\frac{\partial}{\partial\mu}\log[\pi(\mu)]\right)^\top;$$

здесь $\mathbf{E}$ – усреднение по совместной плотности распределения вектора (Y,μ), т.е

$$\mathbf{E}f(Y,\mu) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^d} f(y,u)p_u(y)\pi(u)\,dy\,du.$$

Неравенство Ван Трисса

Теорема

Предположим, что существуют и конечны фишеровские информации I_Y, I_π.
Пусть $\mathbf{E}\|\mu\|^2 < \infty$, *плотность* $p_\mu(Y)$ *ограничена по* μ *при любом фиксированном* Y *и* $\lim_{\|\mu\|\to\infty} \|\mu\|\pi(\mu) = 0$.
Тогда для любой оценки $\hat{\mu}$ *выполнено неравенство*

$$\mathbf{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)^\top \geq (I_Y + I_\pi)^{-1}.$$

Доказательство. Ограничимся для простоты и наглядности случаем $\mu \in \mathbb{R}^1$.

Основная идея - применить неравенство Коши-Буняковского

$$\left[\mathbf{E}a(Y,\mu)\cdot b(Y,\mu)\right]^2 \le \mathbf{E}a^2(Y,\mu)\cdot \mathbf{E}b^2(Y,\mu),$$

где $a(Y,\mu)$ и $a(Y,\mu)$ — некоторые случайные величины.

Возьмем

$$a(Y,\mu) \stackrel{def}{=} \hat{\mu}(Y) - \mu, \quad b(Y,\mu) \stackrel{def}{=} \frac{d}{d\mu}\log[p_\mu(Y)\pi(\mu)].$$

Тогда, интегрируя по частям

$$\begin{aligned}\mathbf{E}a(Y,\mu)b(Y,\mu) &= \int_{\mathbb{R}^n}\left\{\int_{\mathbb{R}^1}[\hat{\mu}(y)-u]\,d_u[p_u(y)\pi(u)]\right\}dy \\ &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^1} p_u(y)\pi(u)\,dy\,du = 1.\end{aligned}$$

Мы неявно использовали, что

$$\lim_{u\to\pm\infty} up_u(y)\pi(u) = 0.$$

Далее

$$\mathbf{E}b^2(Y,\mu) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\left[\frac{d}{du}\log[p_u(y)\pi(u)]\right]^2 p_u(y)\pi(u)\,dydu$$

$$= \int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\left[\frac{1}{p_u(y)}\frac{dp_u(y)}{du} + \frac{1}{\pi(u)}\frac{d\pi(u)}{du}\right]^2 p_u(y)\pi(u)\,dydu$$

$$I_Y + I_\pi + 2\int_{\mathbb{R}^n}\int_{\mathbb{R}^1}\frac{d\pi(u)}{du}\frac{dp_u(y)}{du}dydu$$

$$= I_Y + I_\pi + 2\int_{\mathbb{R}^1}\left\{\frac{d}{du}\int_{\mathbb{R}^n} p_u(y)\,dy\right\}\frac{d\pi(u)}{du}\,du = I_Y + I_\pi.$$

Это равенство завершает доказательство теоремы (см. неравенство Коши-Буняковского, тождество $\mathbf{E}a(Y,\mu)b(Y,\mu) = 1$ и то, что $\mathbf{E}a^2(Y,\mu) = \mathbf{E}[\hat{\mu}(Y) - \mu]^2$).

Если рапределение (Y, μ) гауссовское, то неравенство Ван Трисса становится равенством.

Неравенство Ван Трисса делит статистикие задачи на две принципиально различных класса:

- $I_Y \gg I_\pi$ — классическая статистика (работают предельные теоремы теории вероятности, высокие скорости сходимости оценок, все разумные оценки эквивалентны, контроль риска оценок не представляет проблемы)
- $I_Y \approx I_\pi$ — непараметрическая статистика: оценки принципиально зависят от функции потерь и априорной информации, концентрации оценок зависят существенным образом от μ, предельные теоремы теряют смысл и т.д.

Пример

Пусть $Y_1, \dots, Y_n$ — независимые, одинаково распределенные случайные величины с ограниченной плотностью $p(x)$, $x \in \mathbb{R}^1$. Рассмотрим две задачи оценивания

- Оценивание

$$\mu = \mathbf{P}\{Y_i \le x\} = \int_{-\infty}^{x} p(u)\, du.$$

В этом случае оценка

$$\hat{\mu}_n(Y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \le x\}$$

является асимптотически $(n \to \infty)$ наилучшей оценкой μ. В частности,

$$\lim_{n\to\infty} \mathbf{E}[\sqrt{n}(\hat{\mu} - \mu)]^2 = \mu(1-\mu).$$

- Оценивание $\mu = p(x)$ (x — некоторая фиксированная точка). Идея построения оценки — продифференцировать эмпирическую функцию распределения

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le x\}.$$

Отсюда

$$\hat{\mu}_n(Y) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - Y_i}{h_n}\right).$$

Выбор окна h_n и ядра $K(\cdot)$ зависит от гладкости плотности $p(\cdot)$ в окрестности x. Скорость сходимости $\hat{\mu}_n(Y)$ к μ также зависит от гладкости плотности $p(\cdot)$ в окрестности x.

Вернемся к линейной модели

$$Y = X\mu + \epsilon.$$

Классическая оценка параметра μ в ней определяется как

$$\hat{\mu}_\circ(Y) = \arg\min_\mu \|Y - X\mu\|^2$$

или как корень уравнения

$$X^\top X\hat{\mu}_\circ = X^\top Y.$$

Ясно, что

$$\hat{\mu}_\circ(Y) = (X^\top X)^{-1}X^\top Y = \mu + (X^\top X)^{-1}X^\top \epsilon.$$

Для статистического анализа этой оценки удобно использовать SVD.

Обозначим e_k и λ_k собственные векторы и собственные числа матрицы $X^\top X$:

$$X^\top X e_k = \lambda_k e_k, \quad k = 1, \ldots, p.$$

(Будем считать для определенности, что $\lambda_1 \geq \ldots \geq \lambda_p > 0$.)
Тогда легко видеть, что векторы

$$e_k^* = \frac{X e_k}{\sqrt{\lambda_k}}$$

ортонормальны

$$\langle e_k^*, e_j^* \rangle = \frac{\langle X e_k, X e_j \rangle}{\sqrt{\lambda_k \lambda_j}} = \frac{\langle X^\top X e_k, e_j \rangle}{\sqrt{\lambda_k \lambda_j}} = \frac{\lambda_k \langle e_k, e_j \rangle}{\sqrt{\lambda_k \lambda_j}}.$$

Рассмотрим два линейных преобразования данных Y

$$\tilde{Y}_k^* = \langle Y, e_k^* \rangle = \langle X\mu, e_k^* \rangle + \sigma\xi_k',$$
$$\tilde{Y}_k = \frac{\langle X^\top Y, e_k \rangle}{\lambda_k} = \langle \mu, e_k \rangle + \frac{\sigma}{\sqrt{\lambda_k}}\xi_k';$$

здесь ξ_k' – гауссовские независимые с.в. $\mathcal{N}(0,1)$.
Поэтому, используя для оценивания $X\mu$ вектор $\tilde{Y}^*$, а для оценивания μ — вектор $\tilde{Y}$, сразу же получаем

$$\mathbf{E}\|\hat{\mu}_\circ - \mu\|^2 = \sigma^2 \sum_{k=1}^{p} \frac{1}{\lambda_k}, \quad \mathbf{E}\|X\hat{\mu}_\circ - X\mu\|^2 = \sigma^2 p.$$

здесь $\sigma^2 = \mathbf{E}\epsilon^2(k)$.
Если число обусловленности $X^\top X$ велико или размерность оцениваемого параметра p велика, то чтобы уменьшить риск, мы должны попытаться использовать априорную информацию о μ.

Метод Тихонова (ridge regression)

Например, мы можем считать, что μ – случайный вектор с независимыми $\mathcal{N}(0, 1/S)$ компонентами. В этом случае байессовская оценка имеет вид

$$\hat{\mu}_S = \arg\min_{\mu} \left\{ \frac{\|Y - X\mu\|^2}{2\sigma^2} + \frac{S}{2}\|\mu\|^2 \right\}.$$

Решение этой задачи легко находится

$$\hat{\mu}_S = [X^\top X + \sigma^2 S I]^{-1} X^\top Y = H_S[X^\top X]\hat{\mu}_\circ,$$

где

$$H_S[X^\top X] = [X^\top X + \sigma^2 S I]^{-1} X^\top X$$

и I — единичная матрица.

Заметим, что для матрицы регуляризации в методе Тихонова справедливо спектральное представление

$$H_S[X^\top X] = \sum_{k=1}^{p} H_S(\lambda_k) e_k e_k^\top,$$

где

$$H_S(\lambda) = \frac{\lambda}{\lambda + \sigma^2 S}.$$

Метод Ландвебера

Этот метод основан на простой идее: решить рекуррентным методом уравнение

$$X^\top X\mu = X^\top Y.$$

Поскольку

$$X^\top Y = \left[X^\top X - \alpha I\right]\mu + \alpha\mu$$

для всех $\alpha > 0$, мы имеем

$$\mu = \left[I - \alpha^{-1}X^\top X\right]\mu + \alpha^{-1}X^\top Y.$$

Поэтому мы можем вычислять корень как

$$\hat{\mu}^{(k)} = \left[I - \alpha^{-1}X^\top X\right]\hat{\mu}^{(k-1)} + \alpha^{-1}X^\top Y.$$

Таким образом мы можем оценивать μ без применения SVD и без решения системы линейных уравнений.

Нетрудно проверить, что этот метод сходится при $\alpha > \lambda_1$ и что его регуляризационная матрица имеет следующий вид:

$$H_k(X^\top X) = I - \left(I - \alpha^{-1} X^\top X\right)^{k+1}.$$

Параметр регуляразации метода Ландвебера $S = 1/k$.
К сожалению, этот метод может сходиться очень медленно, если число обусловленности $X^\top X$ велико. Нетрудно видеть, число итераций имеет порядок

$$k \gtrsim \operatorname{cond}(A) \stackrel{\text{def}}{=} \frac{\lambda(1)}{\lambda(n)}.$$

Действительно, k должно быть таким, что $H_k(\lambda_n) \approx 1$. Или, что эквивалентно

$$1 \gg \left(1 - \frac{\lambda_1}{\alpha} \times \frac{\lambda_n}{\lambda_1}\right)^{k+1} \approx \exp\left[-(k+1)\frac{\lambda_1}{\alpha} \times \frac{\lambda_n}{\lambda_1}\right].$$

Существенно улучшенную сходимость имеет так называемый ν-метод.


Landweber, L. (1951). An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.* **73** 615–624.


Engl, H.W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems. Mathematics and its Applications, 375.* Kluwer Academic Publishers Group. Dordrecht.

Spectral cut-off

Для этого метода

$$H_S(\lambda) = \mathbf{1}\{\lambda \geq S\}.$$

Методы регуляризации Тихонова, Ландвебера и spectral cut-off являются частными случаем широкого класса спектральных методов регуляризации, которые имеют следующий вид:

$$\hat{\mu}_S = H_S[X^\top X]\hat{\mu}_\circ(Y).$$

Функция $H_S(\lambda)$, как правило принимает значения из $[0, 1]$ и такова, что

$$\lim_{\lambda \to 0} H_S(\lambda) = 0, \quad \lim_{S \to 0} H_S(\lambda) = 1.$$

Кроме того, не оговаривая далее этого особо, будем предполагать, что эти регуляризации $H_S(\cdot),\ S \in \mathbb{R}^+$ являются упорядоченными (Kneip (1994)) :

- $0 \leq H_S(\lambda) \leq 1$
- если для некоторых $S_1, S_2, \lambda_0 \in \mathbb{R}^+$

$$H_{S_1}(\lambda_0) > H_{S_2}(\lambda_0),$$

тогда для всех $\lambda \in \mathbb{R}^+$

$$H_{S_1}(\lambda) \geq H_{S_2}(\lambda).$$

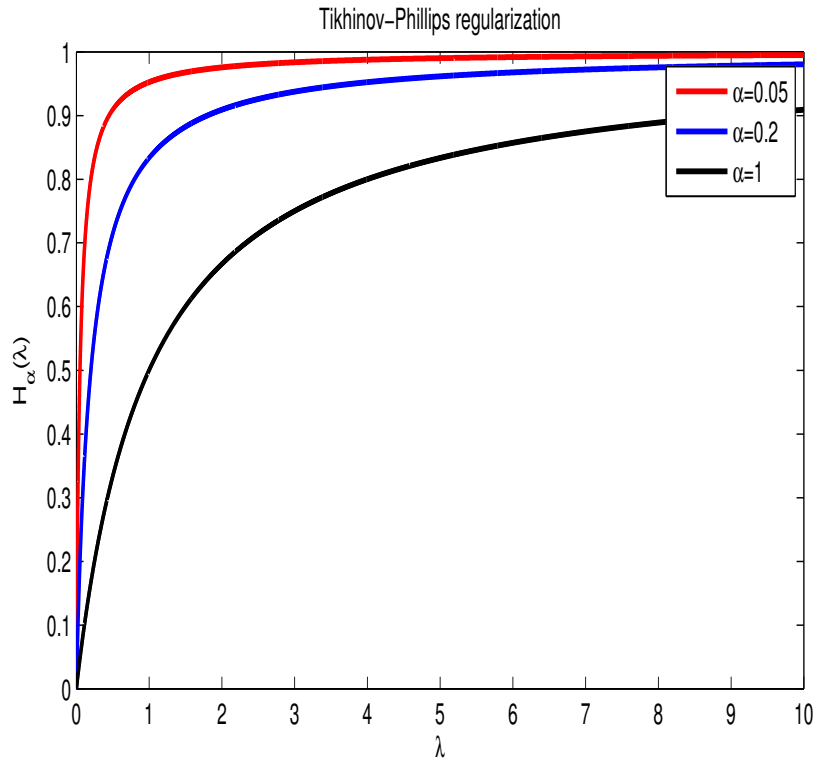
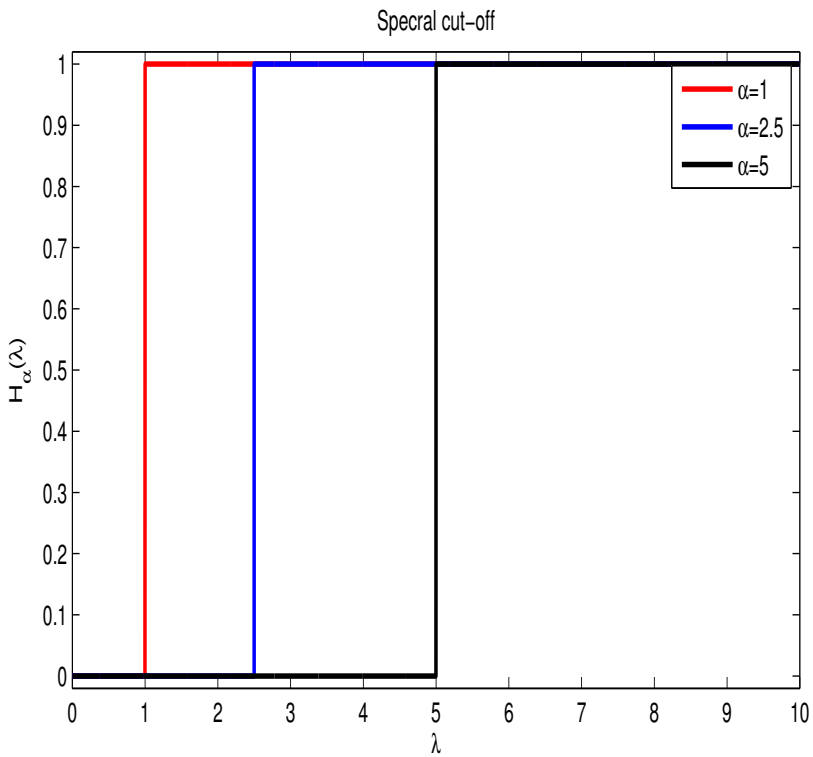


Рис.: $H_S(\lambda) = \lambda/(\sigma^2 S + \lambda)$



$H_S(\lambda) = \mathbf{1}\{\lambda > S\}$

Итак, в нашем распоряжении имеется семейство оценок, индексированных параметром $S \in \mathbb{R}^+$,

$$\hat{\mu}_S = H_S[X^\top X]\hat{\mu}_\circ(Y).$$

Мы хотим найти некоторый метод подбора сглаживающего параметра $\hat{S}^*(Y)$ такой, чтобы

$$\mathbf{E}\|\hat{\mu}_{\hat{S}^*} - \mu\|^2 \leq \mathbf{E}\|\hat{\mu}_{\hat{S}} - \mu\|^2$$

или

$$\mathbf{E}\|X\hat{\mu}_{\hat{S}^*} - X\mu\|^2 \leq \mathbf{E}\|X\hat{\mu}_{\hat{S}} - X\mu\|^2$$

были выполнены для всех μ и любого метода $\hat{S}(Y)$ выбора регуляризирующего параметра.

Если S не зависит от наблюдений, то

$$\mathbf{E}\|\hat{\mu}_S - \mu\|^2 = \sum_{k=1}^{p}[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k},$$

$$\mathbf{E}\|X\hat{\mu}_S - X\mu\|^2 = \sum_{k=1}^{p}\lambda_k[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}H_S^2(\lambda_k).$$

Эти формулы вытекают из спектральных представлений

$$X\hat{\mu}_S(Y) = H_S \cdot \tilde{Y}^*, \quad \hat{\mu}_S(Y) = H_S \cdot \tilde{Y},$$

где

$$\tilde{Y}_k^* = \langle Y, e_k^*\rangle = \langle X\mu, e_k^*\rangle + \sigma\xi_k',$$
$$\tilde{Y}_k = \frac{\langle X^\top Y, e_k\rangle}{\lambda_k} = \langle\mu, e_k\rangle + \frac{\sigma}{\sqrt{\lambda_k}}\xi_k';$$

и того, что $\langle X\mu, e_k^*\rangle = \lambda^{-1/2}\langle X\mu, Xe_k\rangle = \sqrt{\lambda_k}\langle\mu, e_k\rangle$.

Такие представления рисков называются разложениями смещение –дисперсия:

смещения: $\sum_{k=1}^{p}[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2, \quad \sum_{k=1}^{p}\lambda_k[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2$

дисперсии: $\sigma^2\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k}, \quad \sigma^2\sum_{k=1}^{p}H_S^2(\lambda_k).$

Минимаксный подход

Поскольку риски определяютя $\langle \mu, e_k \rangle^2$ и $H_S(\lambda_k)$, то естественно попробовать найти метод регуляризации $H_S(\cdot)$, который их минимизирует. Предположим, что

$$\mu \in \mathcal{M} = \left\{ \mu : \sum_{k=1}^{p} m_k^2 \langle \mu, e_k \rangle^2 \leq E \right\},$$

где $m_1^2 \leq \ldots \leq m_p^2$ — некоторые известные числа. Как правило, $m_k^2 = k^{2s}$.

Нас будут интересовать так называемые минимаксные риски

$$r(\mathcal{M}) = \inf_{H_s} \sup_{\mu \in \mathcal{M}} \mathbf{E} \|\hat{\mu}_S - \mu\|^2,$$

$$r(X\mathcal{M}) = \inf_{H_s} \sup_{\mu \in \mathcal{M}} \mathbf{E} \|X(\hat{\mu}_S - \mu)\|^2.$$

Следующие два результата — часть теоремы Пинскера (1980).

Теорема

$$r(\mathcal{M}) = \sigma^2 \sum_{k=1}^{p} (1 - S|m_k|)_+ \lambda_k^{-1},$$

где $(x)_+ = \max\{x, 0\}$ *и* S *— корень уравнения*

$$\frac{\sigma^2}{S} \sum_{k=1}^{p} (1 - S|m_k|)_+ \lambda_k^{-1} |m_k| = E.$$

При этом минимаксная регуляризация имеет вид

$$H_S(\lambda_k) = (1 - S|m_k|)_+.$$

Теорема

$$r(X\mathcal{M}) = \sum_{k=1}^{p}(1 - S|m_k|\lambda_k^{-1/2})_+,$$

где S *— корень уравнения*

$$\frac{\sigma^2}{S}\sum_{k=1}^{p}(1 - S|m_k|\lambda_k^{-1/2})_+\lambda_k^{-1/2}|m_k| = E.$$

При этом минимакная регуляризация имеет вид

$$H_S(\lambda_k) = (1 - S|m_k|\lambda_k^{-1/2})_+.$$

Доказательство этих результатов вытекает из теоремы о седловой точке т.к. функционалы

$$\mathbf{E}\|\hat{\mu}_S - \mu\|^2 = \sum_{k=1}^{p}[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k},$$

$$\mathbf{E}\|X\hat{\mu}_S - X\mu\|^2 = \sum_{k=1}^{p}\lambda_k[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}H_S^2(\lambda_k).$$

являются линейными по $\langle\mu, e_k\rangle^2$ и квадратичными по $H_S(\lambda_k)$, а мы ищем седловую точку на эллипсоиде

$$\mathcal{M} = \left\{\mu : \sum_{k=1}^{p} m_k^2\langle\mu, e_k\rangle^2 \le E\right\}.$$

Теорема Пинскера

Теорема

Пусть n, $p = \infty$ *и* $\lambda_k \geq Ck^{-q_1}$, $\quad |m_k| \leq Ck^{q_2}$ *для некоторых* q_1, $q_2 > 0$. *Тогда при* $\sigma \to 0$

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{M}} \mathbf{E}\|\hat{\mu} - \mu\|^2 = (1 + o(1))r(\mathcal{M}),$$

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{M}} \mathbf{E}\|X(\hat{\mu} - \mu)\|^2 = (1 + o(1))r(X\mathcal{M});$$

здесь inf вычисляется по всем оценкам вектора μ.

Доказательство. Верхняя граница для минимаксного риска это предыдущая теорема.
Доказательство нижней границы основано на концентрации гауссовской меры на эллипсоиде.
Точнее, предположим, что $\langle \mu, e_k \rangle$ — независимые гауссовские случайные величины с нулевым средним и дисперсиями

$$\Sigma_k^2 = \frac{\sigma^2(1 - S|m_k|)_+}{S\lambda_k|m_k|},$$

где S — корень уравнения

$$\frac{\sigma^2}{S}\sum_{k=1}^{p}(1 - S|m_k|)_+\lambda_k^{-1}|m_k| = E.$$

Тогда

$$\lim_{\sigma\to 0}\mathbf{P}\left\{\sum_{k=1}^{\infty} m_k^2\langle \mu, e_k\rangle^2 \geq (1+\epsilon)E\right\} = 0.$$

Основной недостаток минимаксного подхода это зависимость регуляризации от априорной информации об оцениваемом параметре, в частности, от параметров эллипсоида $\mathcal{M}$.
Несмотря на это, на практике этот подход применяется очень часто. При этом параметры регуляризации выбираются эвристическим методом.
Реальная же ценность минимаксного подхода заключается в асимптотически точных нижних границах.
Идеальным решением был бы адаптивный (т.е. на основе имеющихся наблюдений) подбор параметров регуляризации.

Для решения этой задачи будем рассматривать следующий подход. Пусть у нас имеется семейство оценок

$$\hat{\mu}_S(Y) = H_S(X^\top X)\hat{\mu}_\circ(Y), \quad S \in \mathbb{R}^+.$$

С этим семейством свяжем следующий риск:

$$r(X\mu) \stackrel{\text{def}}{=} \min_S \mathbf{E}\|X\hat{\mu}_S - X\mu\|^2.$$

Наша задача состоит чтобы найти метод $\hat{S}(Y)$ выбора параметра S такой, что

$$\mathbf{E}\|X\mu_{\hat{S}} - X\mu\|^2 \approx r(X\mu)$$

для всех $\mu \in \mathbb{R}^p$.

Теорема Kneip (1994)

Теорема

Пусть

$$\hat{S} = \arg\min_{S} \left\{ \|Y - X\hat{\mu}_S\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) \right\}.$$

Равномерно по $\mu \in \mathbb{R}^p$ *выполняется неравенство*

$$\mathbf{E}\|X(\hat{\mu}_{\hat{S}} - \mu)\|^2 \leq r(X\mu) + K\sigma^2 \sqrt{\frac{r(X\mu)}{\sigma^2}}.$$

Статистический смысл теоремы

1. Предположим, что мы имеем доступ к оракулу, который для любой оценки $\hat{\mu}(Y)$ может предсказать ее риск

$$\mathbf{E}\|X\hat{\mu} - X\mu\|^2.$$

Мы решили для оценивания вектора μ использовать оценки $\hat{\mu}_S(Y)$ и обратиться к оракулу, чтобы выбрать параметр регуляризации S. Тогда

$$r(X\mu) = \min_S \mathbf{E}\|X\hat{\mu}_S - X\mu\|^2$$

можно интерпретировать как риск оракула.

Теорема Кнайпа связывает риск некоторого адаптивного метода выбора и риск оракула, поэтому она часто называется оракульным неравенством.
Можно, естественно, интерпретировать эту теорему как неравенство концентрации, которое описывает концентрацию $X\hat{\mu}_{\hat{S}}$ вблизи $X\mu$ и риска $X\hat{\mu}_{\hat{S}}$ вблизи риска оракула.

Рассмотрим два типичных со статистической точки зрения случая:

- $r(X\mu) \approx \sigma^2$ (модель низкой размерности)

$$\mathbf{E}\|X\hat{\mu}_{\hat{S}} - X\mu\|^2 \approx Kr(X\mu)$$

- $r(X\mu) \gg \sigma^2$ (модель высокой размерности)

$$\mathbf{E}\|X\hat{\mu}_{\hat{S}} - X\mu\|^2 \approx r(X\mu),$$

то есть наша оценка адаптивна.

2. Метод выбора сглаживающего параметра S основанный на принципе несмещенного оценивания риска, по-видимому, впервые был предложен Акаике в 1973 году.

 *Akaike H.* Information theory and an extension of the maximum likelihood principle // Proc. 2nd Intern. Symp. Inf. Theory, 1973, C. 267–281.

 *Kneip A.* Ordered linear smoothers // Annals of Stat. 1994, vol. 22, pp. 835–866.

В рассматриваемой задаче мы выбираем параметр регуляризации

$$\hat{S} = \arg\min_S \left\{ \|Y - X\hat{\mu}_S\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) \right\}$$

в силу того, что для любого фиксированного S

$$\mathbf{E}\left\{ \|Y - X\hat{\mu}_S\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) - \|Y - X\hat{\mu}_\circ\|^2 \right\} = \mathbf{E}\|X\mu - X\hat{\mu}_S\|^2.$$

На самом деле, чтобы обосновать принцип несмещенного оценивания риска нужно гораздо более сильное утверждение, а именно,

$$\|Y - X\hat{\mu}_{\tilde{S}}\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_{\tilde{S}}(\lambda_k) - \|Y - X\hat{\mu}_\circ\|^2 \approx \|X\mu - X\hat{\mu}_{\tilde{S}}\|^2$$

для любого $\tilde{S}(Y)$, зависящего от наблюдений.

Доказательство

1. Посмотрим как связаны и устроены с вероятностной точки зрения следующие расстояния

$$\|X\hat{\mu}_S(Y) - X\mu\|^2 \quad \|X\hat{\mu}_S(Y) - Y\|^2$$

поскольку именно они входят в выражения для риска и эмпирического риска.
Используем спектральное представление наблюдений Y и оценки

$$X\hat{\mu}_S(Y) = XH[X^\top X](X^\top X)^{-1}X^\top Y.$$

Как и ранее e_k и λ_k – собственные векторы и собственные числа $X^\top X$ и

$$e_k^* = \frac{Xe_k}{\sqrt{\lambda_k}}, k = 1, \ldots, n.$$

Разлагая вектор Y по ортонормальной системе e_k^*, имеем

$$\tilde{Y}_k^* = \langle Y, e_k^* \rangle = \langle X\mu, e_k^* \rangle + \sigma\xi_k,$$
$$\tilde{Z}_k = \langle X\hat{\mu}_S(Y), e_k^* \rangle = H_S(\lambda_k)\tilde{Y}_k^*;$$

здесь ξ_k – независимые гауссовские случайные величины с параметрами $(0, 1)$. Обозначим для краткости

$$r_S(X\mu) \stackrel{\text{def}}{=} \mathbf{E}\|X(\hat{\mu}_S - \mu)\|^2$$
$$\bar{r}_S(Y) \stackrel{\text{def}}{=} \|X\hat{\mu}_S - Y\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) - \|X\hat{\mu}_\circ - Y\|^2.$$

Поэтому

$$\|X(\hat{\mu}_S - \mu)\|^2 - r_S(X\mu)$$
$$= -2\sigma \sum_{k=1}^{p} \big[1 - H_S(\lambda_k)\big] \langle X\mu, e_k^* \rangle \xi_k + \sigma^2 \sum_{k=1}^{p} H_S^2(\lambda_k)(\xi_k^2 - 1).$$

и

$$\bar{r}_S(Y) - r_S(X\mu) = \sigma^2 \sum_{k=1}^{p} \big[H_S^2(\lambda_k) - 2H_S(\lambda_k)\big](\xi_k^2 - 1).$$

Напомним, что мы выбираем

$$\hat{S} = \arg\min_{S} \bar{r}_S(Y),$$

где

$$\bar{r}_S(Y) \stackrel{\text{def}}{=} \|X\hat{\mu}_S - Y\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) - \|X\hat{\mu}_\circ - Y\|^2$$

и хотим оценить

$$\mathbf{E}\|X\hat{\mu}_{\hat{S}}(Y) - X\mu\|^2.$$

Поэтому наша задача состоит в том, чтобы построить верхние границы для стохастических слагаемых в правых частях тождеств на предыдущем слайде.

Lemma

Пусть $H_S(\lambda)$ – семейство упорядоченных функций. Тогда для любого $\tilde{S}(Y)$

$$\mathbf{E}|\zeta(\tilde{S})| \le K\sqrt{\mathbf{E}V_\zeta^2(\tilde{S})},$$

где $V_\zeta^2(S) = \mathbf{E}\zeta^2(S)$, а $\zeta(S)$ любой из процессов

$$\zeta(S) = \sum_{k=1}^{p}\big[H_S^2(\lambda_k) - 2H_S(\lambda_k)\big](\xi_k^2 - 1),$$

$$\zeta(S) = \sum_{k=1}^{p} H_S^2(\lambda_k)(\xi_k^2 - 1),$$

$$\zeta(S) = \sum_{k=1}^{p}\big[1 - H_S(\lambda_k)\big]\langle X\mu, e_k^*\rangle\xi_k.$$

С помощью этой леммы мы можем завершить доказательство теоремы. Из тождества

$$\bar{r}_S(Y) - r_S(X\mu) = \sigma^2 \sum_{k=1}^{p} \big[H_S^2(\lambda_k) - 2H_S(\lambda_k)\big](\xi_k^2 - 1)$$

и леммы получаем

$$\begin{aligned}
\mathbf{E} r_{\hat{S}}(X\mu) &\le \mathbf{E}\bar{r}_{\hat{S}}(Y) + K\sigma\sqrt{\mathbf{E} r_{\hat{S}}(X\mu)} \\
&= \mathbf{E}\min_S \bar{r}_S(Y) + K\sigma\sqrt{\mathbf{E} r_{\hat{S}}(X\mu)} \\
&\le \min_S \mathbf{E}\bar{r}_S(Y) + K\sigma\sqrt{\mathbf{E} r_{\hat{S}}(X\mu)} \\
&= r(X\mu) + K\sigma\sqrt{\mathbf{E} r_{\hat{S}}(X\mu)}
\end{aligned}$$

Решая это неравенство относительно $\mathbf{E}r_{\hat{S}}(X\mu)$ и учитывая, что $r(X\mu) \geq \sigma^2$, находим

$$\mathbf{E}r_{\hat{S}}(X\mu) \leq r(X\mu) + K\sigma\sqrt{r(X\mu)}. \tag{1}$$

Далее, для оценивания $\mathbf{E}\|X(\hat{\mu}_{\hat{S}} - \mu)\|^2$ применяем лемму и пользуемся

$$\begin{aligned}
&\|X\hat{\mu}_S - X\mu\|^2 - r_S(X\mu) \\
&= -2\sigma \sum_{k=1}^{p} \big[1 - H_S(\lambda_k)\big] \langle X\mu, e_k^* \rangle \xi_k + \sigma^2 \sum_{k=1}^{p} H_S^2(\lambda_k)(\xi_k^2 - 1).
\end{aligned}$$

Тогда имеем

$$\mathbf{E}\|X\hat{\mu}_{\hat{S}} - X\mu\|^2 \leq \mathbf{E}r_{\hat{S}}(X\mu) + K\sigma\sqrt{\mathbf{E}r_{\hat{S}}(X\mu)}.$$

Подставляя в это неравенство (1), завершаем доказательство.

Доказательство леммы

Пусть $\zeta(S)$ – сепарабельный случайный процесс на $\mathbb{R}^+$.
Пусть также $V(S),\ S \geq 0$ – некоторая положительная, непрерывная, строго возрастающая функция $V(0) = 0$.
Мы начнем с экспоненциальной границы для колмогоровского chaining'а.

Lemma

Пусть $V(x),\ x \in \mathbb{R}^+$ *— непрерывная, неубывающая функция такая, что для некоторого* $\lambda > 0$,

$$\max_{|z|\le\lambda} \max_{0<u<s\le t} \log \mathbf{E} \exp\left\{ z \frac{\sqrt{2}}{\sqrt{2}-1} \frac{\zeta(u)-\zeta(s)}{\sqrt{|V^2(s)-V^2(u)|}} \right\} \le \le Q(\lambda) < \infty, \tag{2}$$

то

$$\log \mathbf{E} \exp\left\{ \lambda \max_{0<u<v\le t} \frac{\zeta(u)-\zeta(s)}{V(t)} \right\} \le \frac{\log(2)\sqrt{2}}{\sqrt{2}-1} + Q(\lambda).$$

Гауссовские процессы

Пусть $\zeta(S)$ – гауссовский процесс с нулевым средним. Возьмем

$$V^2(S) = \mathbf{E}\zeta^2(S).$$

Предположим, что существует постоянная C такая, что

$$\mathbf{E}[\zeta(S_1) - \zeta(S_2)]^2 \leq C|V^2(S_1) - V^2(S_2)|.$$

Тогда $M(S^\circ) = \sup_{S \leq S^\circ} \zeta(S)$ — субгауссовская с.в.

$$\mathbf{E}\exp[\lambda M(S^\circ)] \leq K\exp[K\lambda^2\sigma^2(S^\circ)].$$

Lemma

Если для случайного процесса $\zeta(S)$, $S \in \mathbb{R}^+$ выполнено условие (2), то для любого случайного момента $\tilde{S}$

$$\mathbf{E}\zeta(\tilde{S}) \le K\sqrt{\mathbf{E}V^2(\tilde{S})}.$$

Доказательство. Требуемое неравенство вытекает из следующего факта: для любого $\alpha > 0$

$$\mathbf{E}\max_{S\ge 0}\left\{\zeta(S) - \alpha V^2(S)\right\} \le K\alpha^{-1}. \tag{3}$$

Действительно, если выполнено это неравенство, то

$$\mathbf{E}\zeta(\tilde{S}) \le \alpha\mathbf{E}V^2(\tilde{S}) + K\alpha^{-1}.$$

Поэтому, минимизируя правую часть по α, приходим к требуемому результату.

Для доказательства неравенства (3)

$$\mathbf{E}\max_{S\geq 0}\{\zeta(S)-\alpha V^2(S)\}\leq K\alpha^{-1}$$

введем моменты $S_k^\alpha,\ k=0,1,\ldots$ следующим образом:

$$V^2(S_k^\alpha)=2^{k-1}\alpha^{-1}.$$

Воспользуемся простым неравенством, считая, что $S_{-1}^\alpha=0$

$$\mathbf{E}\max_{S\geq 0}\{\zeta(S)-\alpha V^2(S)\}\leq \mathbf{E}\sum_{k=0}^{\infty}\max_{S_{k-1}^\alpha\leq S\leq S_k^\alpha}\{\zeta(S)-\alpha V^2(S_{k-1}^\alpha)\}_+$$

$$\leq\sum_{k=0}^{\infty}\mathbf{E}\{\max_{S_{k-1}^\alpha\leq S\leq S_k^\alpha}\zeta(S)-\alpha V^2(S_{k-1}^\alpha)\}_+$$

Воспользуемся далее выпуклостью функции $\exp(x)$, точнее, неравенством

$$\{x - A\}_+ \le z^{-1} \exp[z(x - A - 1)],\ z > 0.$$

Поэтому для любой случайной величины ζ получаем (неравенство Чернова)

$$\mathbf{E}\{\zeta - A\}_+ \le z^{-1}\mathbf{E} \exp[z(\zeta - A - 1)].$$

Применим это неравенство для подсчета

$$\mathbf{E}\Big\{ \max_{S_{k-1}^\alpha \le S \le S_k^\alpha} \xi(S) - \alpha V^2(S_{k-1}^\alpha)\Big\}_+$$

положив

$$A = \alpha V^2(S_{k-1}^\alpha), \quad z = \lambda / V(S_k^\alpha).$$

Тогда мы находим, используя chaining-лемму,

$$\mathbf{E}\Big\{\max_{S_{k-1}^{\alpha}\le S\le S_k^{\alpha}}\zeta(S)-\alpha V^2(S_{k-1}^{\alpha})\Big\}_+ \le K\sigma(S_k^{\alpha})\exp[-K\alpha V(S_k^{\alpha})]$$
$$\le K\alpha^{-1}2^{k-1}\exp\big(-K2^{k-1}\big).$$

Таким образом, мы завершаем доказательство леммы, замечая, что

$$\mathbf{E}\max_{S\ge 0}\big\{\zeta(S)-\alpha V^2(S)\big\} \le \sum_{k=0}^{\infty}\mathbf{E}\Big\{\max_{S_{k-1}^{\alpha}\le S\le S_k^{\alpha}}\zeta(S)-\alpha V^2(S_{k-1}^{\alpha})\Big\}_+$$
$$\le K\alpha^{-1}.$$

Завершает наши доказательства следующее наблюдение: любой из процессов

$$\zeta(S) = \sigma^2 \sum_{k=1}^{p} \big[H_S^2(\lambda_k) - 2H_S(\lambda_k)\big](\xi_k^2 - 1),$$

$$\zeta(S) = \sigma^2 \sum_{k=1}^{p} H_S^2(\lambda_k)(\xi_k^2 - 1),$$

$$\zeta(S) = \sum_{k=1}^{p} \big[1 - H_S(\lambda_k)\big] \langle X\mu, e_k^* \rangle \xi_k.$$

удовлетворяет условию

$$\log \mathbf{E} \exp\left\{ \lambda \frac{\zeta(u) - \zeta(s)}{\sqrt{|V^2(s) - V^2(u)|}} \right\} \le K\lambda^2$$

при достаточно малом λ и $V^2(S) = \mathbf{E}\zeta^2(S)$.

Оценивание μ

Что касается оценивания μ, то ситуация оказывается существенно сложнее.
На практике в большинстве случаев используется оценка

$$\hat{\mu}_{\hat{S}}(Y) = H_{\hat{S}}(X^\top X)\hat{\mu}_\circ(Y),$$

$$\text{где} \quad \hat{S} = \arg\min_S \left\{ \|Y - X\hat{\mu}_S\|^2 + 2\sigma^2 \sum_{k=1}^{p} H_S(\lambda_k) \right\}.$$

К сожалению, для этого метода не известны нетривиальные верхние границы для $\mathbf{E}\|\hat{\mu}_{\hat{S}}(Y) - \mu\|^2$.
Конечно, можно воспользоваться следующим неравенством:

$$\begin{aligned}\mathbf{E}\|\hat{\mu}_{\hat{S}}(Y) - \mu\|^2 &= \mathbf{E}\|(X^\top X)^{-1}(X^\top X)(\hat{\mu}_{\hat{S}}(Y) - \mu)\|^2 \\ &\le \|(X^\top X)^{-1}X^\top\|^2 \mathbf{E}\|X\hat{\mu}_{\hat{S}}(Y) - X\mu\|^2,\end{aligned}$$

но это завышенная оценка, для матриц с плохой обусловленностью.

В принципе, основная идея выбора сглаживающего параметра остается близкой к методу несмещенного оценивания риска. Однако ее реализация оказывается специфической.
Поскольку мы хотим теперь оценивать параметр μ, а не $X\mu$ как ранее, введем новый эмпирический риск

$$R_S[Y, Pen] = \|\hat{\mu}_\circ(Y) - \hat{\mu}_S\|^2 + Pen(S) - \|\hat{\mu}_\circ - \mu\|^2$$

пенализованный некоторой функцией $Pen(S)$. Напомним, что ранее мы использовали

$$R_S[Y, Pen] = \|X\hat{\mu}_\circ(Y) - X\hat{\mu}_S\|^2 + Pen(S) - \|X\hat{\mu}_\circ - X\mu\|^2$$

Разница между этими методами объясняется тем, что для восстановления μ и $X\mu$ используются по сути разные, наблюдения

$$\tilde{Y}_k^* = \langle X\mu, e_k^* \rangle + \sigma\xi_k, \quad \text{для оценки } X\mu$$
$$\tilde{Y}_k = \langle \mu, e_k \rangle + \frac{\sigma}{\sqrt{\lambda_k}}\xi_k \quad \text{для оценки } \mu.$$

Поскольку отношение λ_1/λ_p может быть в принципе очень большим, вторая модель содержит существенно "больше" шума, чем первая. Этот факт существенно меняет метод выбора параметра регуляризации.

Как и ранее мы оцениваем μ

$$\hat{\mu}_S = H_S \cdot \tilde{Y}$$

и выбираем параметр регуляризации

$$\hat{S} = \arg\min_S \left\{ \|\tilde{Y} - \hat{\mu}_S\|^2 + Pen(S) \right\}.$$

Наша цель выбрать $Pen(S)$ так, чтобы риск оценки $\hat{\mu}_{\hat{S}}$ как можно лучше приближал риск оракула

$$\min_S r_S(\mu),$$

где

$$r_S(\mu) = \mathbf{E}\|\hat{\mu}_S - \mu\|^2 = \sum_{k=1}^{p} [1 - H_S(\lambda_k)]^2 \langle \mu, e_k \rangle^2 + \sigma^2 \sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k}.$$

Основная идея решения этой задачи очень проста: выберем Pen так, чтобы

$$r_S(\mu) \le R_S[\tilde{Y}, Pen] \stackrel{\text{def}}{=} \|\tilde{Y} - \hat{\mu}_S\|^2 + Pen(S) - \|\tilde{Y} - \hat{\mu}_\circ\|^2.$$

Конечно, найти такую функцию не удастся, но можно найти минимальную функцию $Pen(S)$ такую, что

$$\mathbf{E} \sup_S \Big\{ r_S(\mu) - R_S[\tilde{Y}, Pen] \Big\}_+ \le \sigma^2.$$

Оказывается, что для упорядоченных регуляризаций эта задача имеет решение. Оно связано с поиском "детернинированной оболочки"для случайного процесса

$$\zeta(S) = \sum_{k=1}^{p} \lambda_k^{-1} \Big\{ 2H_S(\lambda_k) - H_S^2(\lambda_k) \Big\} (\xi_k^2 - 1),$$

где ξ_k – независимые $\mathcal{N}(0,1)$.
Задача состоит в том, чтобы найти минимальную детерминированную функцию $\mathcal{E}(S)$ такую, что

$$\mathbf{E} \sup_S \big\{ \zeta(S) - \mathcal{E}(S) \big\}_+ \leq K.$$

Пусть $pen(S)$ корень уравнения

$$\mathbf{E}[\zeta(S) - pen(S)]_+ = 1.$$

Теорема

Для любого $\gamma > 0$

$$\mathbf{E} \sup_S \left\{\zeta(S) - (1+\gamma)pen(S)\right\}_+^{1+\epsilon} \leq \frac{K}{(\gamma - \epsilon)_+^3}.$$

Положим

$$Pen(S) = 2\sigma^2 \sum_{k=1}^{p} \frac{H_S[\lambda_k]}{\lambda_k} + \sigma^2(1+\gamma)pen(S).$$

Теорема

Для любого $\gamma > 0$, равномерно по $\mu \in \mathbb{R}^p$ выполнено нерравенство

$$\mathbf{E}\|\hat{\mu}_{\hat{S}} - \mu\|^2 \le \bar{r}(\mu) + \Psi_\gamma\left(\frac{\sigma^2}{\bar{r}(\mu)}\right)\bar{r}(\mu),$$

где

$$\bar{r}(\mu) = \min_S \left\{ \mathbf{E}\|\hat{\mu}_S - \mu\|^2 + \sigma^2(1+\gamma)pen(S) \right\}$$

и $\Psi_\gamma(\cdot)$ – ограниченная функция такая, что

$$\lim_{x \to 0} \Psi_\gamma(x) = 0.$$

Грубая оценка для функции $pen(S)$ имеет следующий вид:

$$KV(S)\sqrt{\log[V^2(S)]} \le pen(S) \le KV(S)\log[V^2(S)],$$

где

$$V^2(S) = \mathbf{E}\zeta^2(S) = 2\sum_{k=1}^{p} \frac{[2H_S(\lambda_k) - H_S^2(\lambda_k)]^2}{\lambda_k^2} \approx \sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k^2}.$$

Вспомним, что оракул минимизирует по S риск

$$\mathbf{E}\|\hat{\mu}_S - \mu\|^2 = \sum_{k=1}^{p}[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k}$$

в то время как адаптивный метод минимизирует

$$\sum_{k=1}^{p}[1 - H_S(\lambda_k)]^2\langle\mu, e_k\rangle^2 + \sigma^2\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k}$$

$$+\sigma^2\sqrt{\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k^2}\log\left[\sum_{k=1}^{p}\frac{H_S^2(\lambda_k)}{\lambda_k^2}\right]}$$

Если $\lambda_k \approx k^{-\gamma},\ \gamma > 0$, то

$$\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k} \gg \sqrt{\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k^2} \log\left[\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k^2}\right]}$$

и мы имеем аналог теоремы Кнайпа.
Если же $\lambda_k \approx \exp(-\gamma k),\ \gamma > 0$, то

$$\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k} \ll \sqrt{\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k^2} \log\left[\sum_{k=1}^{p} \frac{H_S^2(\lambda_k)}{\lambda_k^2}\right]}$$

и мы не можем приблизиться к риску оракула.