UNIVERSITY OF OXFORD

DEPARTMENT OF STATISTICS

Stein's Method, Stein's Shrinkage Estimator, and Stein's Unbiased Risk Estimate

Gesine Reinert

Department of Statistics
University of Oxford
and
The Alan Turing Institute

UK Easter Probability Meeting 2023

This is joint work with
Max Fathi, Larry Goldstein, and Adrien Saumard

Annals of Statistics, (2022) vol 50, No 4, 2732–2766

https://arxiv.org/pdf/2004.01378.pdf

1. Background: $\mathbb{E}[\langle X - \boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle]$
   - Stein's method
   - Stein's shrinkage estimator
   - Stein's Unbiased Risk Estimate (SURE)

2. Non-Gaussian models
   - Paths to extension
   - General Stein kernels $\mathbb{E}[\langle X - \boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle]$
     - Stein kernel consequence: shrinkage
     - Stein kernel consequence: SURE
   - Zero-biasing $\mathbb{E}[\langle Y, f(Y)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(Y^*)\rangle]$
     - Zero bias consequence: shrinkage
     - Zero bias consequence: SURE

3. What else

Outline

1. Background: $\mathbb{E}[\langle X - \boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle]$
   - Stein's method
   - Stein's shrinkage estimator
   - Stein's Unbiased Risk Estimate (SURE)

2. Non-Gaussian models
   - Paths to extension
   - General Stein kernels $\mathbb{E}[\langle X - \boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle]$
     - Stein kernel consequence: shrinkage
     - Stein kernel consequence: SURE
   - Zero-biasing $\mathbb{E}[\langle Y, f(Y)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(Y^*)\rangle]$
     - Zero bias consequence: shrinkage
     - Zero bias consequence: SURE

3. What else

Three Big Stein Things

- Stein's method: assessing distances between distributions (*Stein 1972*)
- Stein shrinkage: adjust estimators in high dimension (*Stein 1956, James and Stein 1961*)
- Stein's Unbiased Risk Estimate: estimates the risk of the shrinkage estimator (*Stein 1981*)

An underlying key observation

$X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if for all smooth functions f,

$$\mathbb{E}(X-\mu)f(X) = \sigma^2 \mathbb{E} f'(X).$$

$X \sim \mathcal{N}_d(\boldsymbol{\theta}, \Sigma) =: \nu$ if and only if for all $f \in W^{1,2}(\nu)$

$$\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle].$$

Here $\langle A, B\rangle = Tr(AB^T)$. For example, $\langle \mathrm{Id}, \mathrm{Id}\rangle = d$.

The space $W^{1,2}(\nu)$ is a so-called *Sobolev space*, induced by

$$||f||^2_{W^{1,2}(\nu)} := ||f||^2_{L^2(\nu)} + ||\nabla f||^2_{L^2(\nu)}.$$

Stein's method

Aim: assess distance to a normal distribution (or other distribution)

For a random vector W with $\mathbb{E}\mathsf{W} = \mu, \operatorname{Var} \mathsf{W} = \Sigma$, if

$$\mathbb{E}[\langle \mathsf{W} - \boldsymbol{\theta}, \mathsf{f}(\mathsf{W}) \rangle] - \mathbb{E}[\langle \Sigma, \nabla \mathsf{f}(\mathsf{W}) \rangle]$$

is close to zero for many functions f, then the distribution of W should be close to ν in distribution.

Stein's shrinkage estimator

Aim: estimate $\boldsymbol{\theta}$ in $\mathcal{N}_d(\boldsymbol{\theta}, \sigma^2 \, \mathrm{Id})$ from data $X \in \mathbb{R}^d$

For $\lambda \geq 0$ put

$$S_\lambda(X) = X\left(1 - \frac{\lambda}{\|X\|^2}\right)$$

For $d \geq 3$ there exists a range of positive values for λ for which $S_\lambda(X)$ has a strictly smaller mean squared error than $S_0(X)$.

Mean squared error:

$$E_{\boldsymbol{\theta}}||S_\lambda(X) - \boldsymbol{\theta}||^2 = E_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda\left\langle X - \boldsymbol{\theta}, \frac{X}{\|X\|^2}\right\rangle + \frac{\lambda^2}{||X||^2}\right\}$$

Why? For $S_\lambda(X) = X\left(1 - \frac{\lambda}{||X||^2}\right)$: Use $f(x) = -\lambda\frac{x}{||x||^2}$ in

$$\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \sigma^2 \,\mathrm{Id}, \nabla f(X)\rangle]$$

with $\nabla f(x) = -\lambda\frac{1}{||x||^2}\mathrm{Id} + \lambda\frac{2}{||x||^4}xx^T$ to see that

$$-2\lambda\,\mathbb{E}\left\langle X-\boldsymbol{\theta}, \frac{X}{||X||^2}\right\rangle = -2\sigma^2 d\mathbb{E}\frac{\lambda}{||X||^2} + 2\sigma^2\mathbb{E}\frac{\lambda}{||X||^2}$$

and so

$$E_{\boldsymbol{\theta}}||S_\lambda(X) - \boldsymbol{\theta}||^2 = E_{\boldsymbol{\theta}}\left\{||X-\boldsymbol{\theta}||^2 + \frac{\lambda^2}{||X||^2}\left[-2\sigma^2(d-2)+\lambda\right]\right\}.$$

The last contribution is negative when $\lambda < 2\sigma^2(d-2)$.

Stein's Unbiased Risk Estimate (SURE)

Aim: Unbiased estimator for the mean squared error (risk) of an estimator of $\boldsymbol{\theta}$ in $\mathcal{N}_d(\boldsymbol{\theta}, \Sigma)$ from data $X \in \mathbb{R}^d$

Consider an estimator for $\boldsymbol{\theta}$ of the form

$$S(\mathrm{X}) = \mathrm{X} + \mathrm{f}(\mathrm{X}).$$

Then

$$\mathrm{SURE}(\mathrm{f}, \mathrm{X}) := \mathrm{Tr}(\Sigma) + \|\mathrm{f}(\mathrm{X})\|^2 + 2\sum_{i,j=1}^{d} \sigma_{ij}\partial_j f_i(\mathrm{X})$$

is unbiased for $\mathbb{E}\|S(\mathrm{X}) - \boldsymbol{\theta}\|^2$.

Why $\mathrm{SURE}(f, X) := \mathrm{Tr}(\Sigma) + \|f(X)\|^2 + 2\langle \Sigma, \nabla f(X)\rangle$?

$$\begin{aligned} \|S(X) - \boldsymbol{\theta}\|^2 &= \|X - \boldsymbol{\theta} + f(X)\|^2 \\ &= \|X - \boldsymbol{\theta}\|^2 + \|f(X)\|^2 + 2\langle f(X), X - \boldsymbol{\theta}\rangle. \end{aligned}$$

An unbiased estimator for $\|X - \boldsymbol{\theta}\|^2$ is $Tr(\Sigma) = d\sigma^2$.

An unbiased estimator for $\|f(X)\|^2$ is $\|f(X)\|^2$ as $\boldsymbol{\theta}$ does not appear.

Taking expectations and using $\mathbb{E}[\langle X - \boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle]$ eliminates θ: an unbiased estimator for $2\langle f(X), X - \boldsymbol{\theta}\rangle$ is

$$2\sum_{i,j=1}^{d} \sigma_{ij}\partial_j f_i(X) = 2\langle \Sigma, \nabla f(X)\rangle.$$

Outline

1 Background: $\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle]$
- Stein's method
- Stein's shrinkage estimator
- Stein's Unbiased Risk Estimate (SURE)

2 Non-Gaussian models
- Paths to extension
- General Stein kernels $\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle]$
  - Stein kernel consequence: shrinkage
  - Stein kernel consequence: SURE
- Zero-biasing $\mathbb{E}[\langle Y, f(Y)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(Y^*)\rangle]$
  - Zero bias consequence: shrinkage
  - Zero bias consequence: SURE

3 What else

Stein characterisations for non-Gaussian models

The important equation above is

$$\mathbb{E}[\langle \mathsf{X} - \boldsymbol{\theta}, \mathsf{f}(\mathsf{X})\rangle] = \mathbb{E}[\langle \Sigma, \nabla \mathsf{f}(\mathsf{X})\rangle]$$

which characterises the multivariate normal distribution.

Characterisations for other distributions are available! (E.g. *Mijoule, R., Swan 2022*)

Hence we can extend Stein's shrinkage estimator and SURE to non-Gaussian models.

Some special cases which assume some symmetry: *Cellier et al. 1989, Srivastava and Bilodeau 1989, Evans and Stark 1996, Chen, Wiesel and Hero 2011, Fourdrinier et al. 2018*. Here no such assumption is needed.

Shrinkage

shrinkage estimator for $\lambda \geq 0$

$$S_\lambda(\mathsf{X}) = \mathsf{X}\left(1 - \frac{\lambda}{||\mathsf{X}||^2}\right)$$

has mean squared error

$$\mathbb{E}_{\boldsymbol{\theta}}||S_\lambda(\mathsf{X}) - \boldsymbol{\theta}||^2 = \mathbb{E}_{\boldsymbol{\theta}}\left\{||\mathsf{X} - \boldsymbol{\theta}||^2 - 2\lambda\left\langle \mathsf{X} - \boldsymbol{\theta}, \frac{\mathsf{X}}{||\mathsf{X}||^2}\right\rangle + \frac{\lambda^2}{||\mathsf{X}||^2}\right\}.$$

We cannot do anything about $\mathbb{E}_{\boldsymbol{\theta}}||\mathsf{X} - \boldsymbol{\theta}||^2$. The middle, red term is of the form

$$\mathbb{E}_{\boldsymbol{\theta}}\langle \mathsf{X} - \boldsymbol{\theta}, f(\mathsf{X})\rangle.$$

SURE

SURE for $S(\mathrm{X}) = \mathrm{X} + \mathrm{f}(\mathrm{X})$: unbiased estimate of the expectation of

$$\begin{aligned} \|S(\mathrm{X}) - \boldsymbol{\theta}\|^2 &= \|\mathrm{X} - \boldsymbol{\theta} + \mathrm{f}(\mathrm{X})\|^2 \\ &= \|\mathrm{X} - \boldsymbol{\theta}\|^2 + \|\mathrm{f}(\mathrm{X})\|^2 + 2\,\langle \mathrm{f}(\mathrm{X}), \mathrm{X} - \boldsymbol{\theta}\rangle. \end{aligned}$$

An unbiased estimator for $\|\mathrm{X} - \boldsymbol{\theta}\|^2$ is $Tr(\Sigma) = d\sigma^2$.

An unbiased estimator for $\|\mathrm{f}(\mathrm{X})\|^2$ is $\|\mathrm{f}(\mathrm{X})\|^2$.

The expectation of the last, red term is of the form

$$\mathbb{E}_{\boldsymbol{\theta}}\langle \mathrm{X} - \boldsymbol{\theta}, f(\mathrm{X})\rangle.$$

Path to extension 1: Stein kernels

In dimension 1: X has law $\mathcal{N}(\theta, \sigma^2) \iff$ for all smooth f

$$\mathbb{E}[(X-\theta)f(X)] = \sigma^2 \mathbb{E}[f'(X)].$$

A *Stein kernel* $T_{X-\theta}$ for the distribution of a mean θ random variable X is a random variable for which for all smooth f

$$\mathbb{E}[(X-\theta)f(X)] = \mathbb{E}[T_{X-\theta} f'(X)].$$

(*Cacoullos and Papathanasiou 92*).

For $\mathcal{N}(\theta, \sigma^2)$, $T_{X-\theta} = \sigma^2$ does not depend on θ. In general, it does.

The density of a Stein kernel in 1 dim is explicit: If X has pdf p_X, mean zero, variance σ^2, then $T = T(X)$ can be chosen to be

$$T(X) = p_X(X)^{-1} \int_{-\infty}^{X} y\, p_X(y)\, dy$$

Check:

$$\begin{aligned} \mathbb{E}[Tf'(X)] &= \int_{-\infty}^{\infty} f'(x) p_X(x)^{-1} \int_{-\infty}^{x} y\, p_X(y)\, dy\, p_X(x)\, dx \\ &= \int_{-\infty}^{\infty} f'(x) \int_{-\infty}^{x} y\, p_X(y)\, dy\, dx. \end{aligned}$$

Now (assuming interchanging integrals is allowed)

$$\begin{aligned}
&\int_0^\infty f'(x) \int_x^\infty y p_X(y)\, dy\, dx \\
&\quad = \int_0^\infty y\, p_X(y) \int_0^y f'(x)\, dx\, dy \\
&\quad = \int_0^\infty y\, p_X(y)[f(y) - f(0)]\, dy \\
&\quad = \mathbb{E}[X\, f(X) \mathbb{1}(X \geq 0)] - f(0)\mathbb{E}[X \mathbb{1}(X \geq 0)];
\end{aligned}$$

similarly for the other integral. As $\mathbb{E}[X] = 0$,

$$\mathbb{E}[T\, f'(X)] = \mathbb{E}[X\, f(X)].$$

Path to extension 2: Zero-bias couplings

In dimension 1: X has law $\mathcal{N}(\theta, \sigma^2) \iff$ for all smooth f

$$\mathbb{E}[(X-\theta)f(X)] = \sigma^2 \mathbb{E}[f'(X)].$$

A random variable X^* has the *zero bias distribution* for the distribution of a mean 0, variance σ^2 random variable X if for all smooth f

$$\mathbb{E}[Xf(X)] = \sigma^2 \mathbb{E}[f'(X^*)].$$

(*Goldstein and R. 97*). For X with mean θ

$$X^* := (X-\theta)^* + \theta.$$

For $\mathcal{N}(\theta, \sigma^2)$, $X^* = X$ in distribution.

The density of the zero bias distribution in 1 dim is explicit: If X has mean zero, distribution μ, variance σ^2, then X^* is continuous with density

$$p^*(x) = \frac{1}{\sigma^2}\mathbb{E}[X\mathbb{1}(X \geq x)].$$

Check:

$$\begin{aligned} \sigma^2\mathbb{E}[f'(X^*)] &= \int_{-\infty}^{\infty} f'(x)\mathbb{E}[X\mathbb{1}(X \geq x)]\,dx \\ &= \int_{-\infty}^{\infty} f'(x) \int_{x}^{\infty} y\mathbb{1}(y \geq x)\,d\mu(y)\,dx. \end{aligned}$$

Now (assuming interchanging integrals is allowed)

$$\begin{aligned}
&\int_0^\infty f'(x) \int_{-\infty}^\infty y \mathbb{1}(y \geq x)\, d\mu(y)\, dx \\
&= \int_0^\infty y \int_0^\infty f'(x)\, dx\, d\mu(y) \\
&= \int_0^\infty y p_X(y)[f(y) - f(0)]\, dy \\
&= \mathbb{E}[Xf(X)\mathbb{1}(X \geq 0)] - f(0)\mathbb{E}[X\mathbb{1}(X \geq 0)];
\end{aligned}$$

similarly for the other integral. As $\mathbb{E}[X] = 0$,

$$\sigma^2 \mathbb{E}[f'(X^*)] = \mathbb{E}[Xf(X)].$$

Mixture construction

Let $Y_j, j = 1, \ldots, n$ be independent, mean zero $\mathbb{R}$ valued random vectors with variances σ_j^2 and associated zero bias variables Y_j^*.

Then $Y = \sum_{j=1}^n Y_j$has zero bias variable

$$Y^* = Y - Y_I + Y_I^*$$

where I is independent of the $Y_j's$ and $\mathbb{P}(I = j) = \frac{\sigma_j^2}{\sigma^2}$.

When $Y = Y_j$ with probability $\mu(j)$, then

$$Y^* = Y_J^*$$

where J is independent of the $Y_j's$ and $\mathbb{P}(J = j) = \frac{\sigma_j^2}{\sigma^2}\mu(j)$.

Stein kernels

Chatterjee 2008; Nourdin and Peccati 2012; Mijoule, R. Swan 2022

Given a random vector $\mathrm{X} \in \mathbb{R}^d$ with mean $\boldsymbol{\theta}$ and distribution ν which is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, a Stein kernel $T_{\mathrm{X}-\boldsymbol{\theta}}$ for $\mathrm{X} - \boldsymbol{\theta}$ is a matrix-valued function such that for all $\mathrm{f} \in W^{1,2}(\nu)$

$$\mathbb{E}[\langle \mathrm{X} - \boldsymbol{\theta}, \mathrm{f}(\mathrm{X})\rangle] = \mathbb{E}[\langle T_{\mathrm{X}-\boldsymbol{\theta}}, \nabla \mathrm{f}(\mathrm{X})\rangle].$$

Using $\mathrm{f}(\mathrm{x}) = \mathrm{x}$ we obtain $\mathbb{E}[T_{\mathrm{X}-\boldsymbol{\theta}}] = \Sigma$, the covariance matrix.

For multivariate normal, Σ can serve as Stein kernel.

Stein kernel consequence: shrinkage

Let X have mean $\boldsymbol{\theta}$, positive semidefinite Σ with largest eigenvalue κ, Stein kernel T, and consider $S_\lambda(X) = X\left(1 - \frac{\lambda}{||X||^2}\right)$. Then with $f(x) = x/||x||^2$

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}||S_\lambda(X) - \boldsymbol{\theta}||^2 &= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda\left\langle X - \boldsymbol{\theta}, \frac{X}{||X||^2}\right\rangle + \frac{\lambda^2}{||X||^2}\right\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda\left\langle X - \boldsymbol{\theta}, f(X)\right\rangle + \frac{\lambda^2}{||X||^2}\right\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda\left\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\right\rangle + \frac{\lambda^2}{||X||^2}\right\}.
\end{aligned}
$$

Now for $\mathbb{E}_{\boldsymbol{\theta}}\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle$, with $f(x) = x/||x||^2$,

$$\mathbb{E}_{\boldsymbol{\theta}}\langle T_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle = \mathbb{E}_{\boldsymbol{\theta}}[\langle \Sigma, \nabla f(X)\rangle] + \mathbb{E}_{\boldsymbol{\theta}}[\langle T_{X-\boldsymbol{\theta}} - \Sigma, \nabla f(X)\rangle].$$

As

$$\nabla f(x) = \frac{1}{\|x\|^2}\mathrm{Id} - \frac{2}{\|x\|^4}xx^T, \text{ we have}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[\langle \Sigma, \nabla f(X)\rangle] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\langle \Sigma, \frac{1}{\|X\|^2}\mathrm{Id} - \frac{2}{\|X\|^4}XX^T\rangle\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathrm{Tr}(\Sigma)}{\|X\|^2} - \frac{2\,\mathrm{Tr}(\Sigma XX^T)}{\|X\|^4}\right] \\
&\geq \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathrm{Tr}(\Sigma) - 2\kappa}{\|X\|^2}\right].
\end{aligned}$$

Similarly, (omitting the suffix $X-\boldsymbol{\theta}$ in T)

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[\langle T-\Sigma, \nabla f(X)\rangle] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\langle T-\Sigma, \frac{1}{\|X\|^2}\,\mathrm{Id} - \frac{2}{\|X\|^4}XX^{T}\rangle\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathrm{Tr}(T-\Sigma)}{\|X\|^2} - \frac{2\,\mathrm{Tr}(T-\Sigma XX^{T})}{\|X\|^4}\right].\end{aligned}$$

Using the Cauchy-Schwarz inequality and $\mathbb{E}_{\boldsymbol{\theta}}[T]=\Sigma$,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathrm{Tr}(T-\Sigma)}{\|X\|^2}\right] \le \sqrt{\mathrm{Var}(\mathrm{Tr}(T))}\sqrt{\mathbb{E}[\|X\|^{-4}]}.$$

For the second term we also use Cauchy Schwarz, after bounding

$$|\,\mathrm{Tr}((T-\Sigma)XX^{T})| = |\langle T-\Sigma, XX^{T}\rangle| \le \|X\|^2\|T-\Sigma\|.$$

Result for shrinkage

Let $B_\lambda = \frac{\lambda}{d}\sqrt{E_{\boldsymbol{\theta}}[d^2\|\mathrm{X}\|^{-4}]}\left\{\sqrt{\mathrm{Var}(\mathrm{Tr}(T))} + 2\sqrt{E[\|T-\Sigma\|^2]}\right\}$.
Then

$$\mathbb{E}_{\boldsymbol{\theta}}||S_\lambda(\mathrm{X})-\boldsymbol{\theta}||^2 \leq \mathbb{E}||\mathrm{X}-\boldsymbol{\theta}||^2 + \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\lambda}{||\mathrm{X}||^2}\left(\lambda - 2\left(\mathrm{Tr}(\Sigma) - 2\kappa\right)\right)\right] + 2B_\lambda.$$

If for some d_0: $\sup_{d \geqslant d_0} \mathbb{E}_{\boldsymbol{\theta}}[d^2||\mathrm{X}||^{-4}] < \infty$ and $\mathrm{Var}(\mathrm{Tr}(T)) = o(d^2)$, and $\mathbb{E}[||T-\Sigma||^2] = o(d^2)$ and $\lambda = O(d)$, then $B_\lambda = o(d)$.

In this situation, for $\lambda \in [0, 2(\mathrm{Tr}(\Sigma) - 2\kappa)]$ the risk of S_λ is no larger than that of S_0 asymptotically.

Example: Student distribution

$\mathbf{X} = \mathbf{Y} + \theta$ in $\mathbb{R}^d$ with $\mathbf{Y}$ from the family of multivariate central Student-t distributions, with $k \geq 5$ degrees of freedom, shape given by the identity matrix in $\mathbb{R}^{d \times d}$ and $d = 2m \geq 6$, even.

The covariance matrix is then $\Sigma = \sigma^2 \, \mathrm{Id}$.

Using results from *Mijoule, R., Swan 2022*,

$$T = \left(\frac{\mathbf{Y}^\mathsf{T}\mathbf{Y} + k\sigma^2}{d + k - 2} \right) \mathrm{Id}$$

is a Stein kernel.

The above limiting regime holds as long as $1/k = o(1)$.

Stein kernel consequence: SURE

SURE for $S(\mathrm{X}) = \mathrm{X} + \mathrm{f}(\mathrm{X})$: unbiased estimate of the expectation of

$$\|S(\mathrm{X}) - \boldsymbol{\theta}\|^2 \quad = \quad \|\mathrm{X} - \boldsymbol{\theta}\|^2 + \|\mathrm{f}(\mathrm{X})\|^2 + 2\,\langle \mathrm{f}(\mathrm{X}), \mathrm{X} - \boldsymbol{\theta}\rangle.$$

An unbiased estimator for $\|\mathrm{X} - \boldsymbol{\theta}\|^2$ is $Tr(\Sigma) = d\sigma^2$.

An unbiased estimator for $\|\mathrm{f}(\mathrm{X})\|^2$ is $\|\mathrm{f}(\mathrm{X})\|^2$.

By the Stein kernel, an unbiased estimate of $\langle \mathrm{f}(\mathrm{X}), \mathrm{X} - \boldsymbol{\theta}\rangle$ is

$$\langle T_{\mathrm{X}-\boldsymbol{\theta}}, \nabla \mathrm{f}(\mathrm{X})\rangle.$$

We introduce

$$\mathrm{SURE}_k(\mathrm{f}, \mathrm{X}) = \mathrm{Tr}(\Sigma) + ||\mathrm{f}(\mathrm{X})||^2 + 2\langle T_{\mathrm{X}-\boldsymbol{\theta}}, \nabla \mathrm{f}(\mathrm{X})\rangle.$$

The bias of SURE

In practice it may not be possible to calculate SURE_k. What if we use SURE instead? Recall

$$\mathrm{SURE}(\mathrm{f}, \mathrm{X}) = \mathrm{Tr}(\Sigma) + \|\mathrm{f}(\mathrm{X})\|^2 + 2\langle \Sigma, \nabla \mathrm{f}(\mathrm{X})\rangle$$

The bias is

$$\begin{aligned} \mathrm{Bias}_{\boldsymbol{\theta}}(\mathrm{SURE}(\mathrm{f}, \mathrm{X})) &= \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{SURE}(\mathrm{f}, \mathrm{X}))] - \mathbb{E}_{\boldsymbol{\theta}}\|S(\mathrm{X}) - \boldsymbol{\theta}\|^2 \\ &= \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{SURE}(\mathrm{f}, \mathrm{X}))] - \mathbb{E}[\mathrm{SURE}_k(\mathrm{f}, \mathrm{X})] \\ &= 2\,\mathbb{E}[\langle \Sigma - T_{\mathrm{X}-\boldsymbol{\theta}}, \nabla \mathrm{f}(\mathrm{X})\rangle]. \end{aligned}$$

Thus,

$$|\mathrm{Bias}_{\boldsymbol{\theta}}(\mathrm{SURE}(\mathrm{f}, \mathrm{X}))| \leq 2|\mathbb{E}[\langle \Sigma - T_{\mathrm{X}-\boldsymbol{\theta}}, \nabla \mathrm{f}(\mathrm{X})\rangle]|.$$

Corollary

If $X = Y + \boldsymbol{\theta}$, where Y has covariance matrix Σ and Stein kernel T, and if $f(x) = -\lambda \frac{x}{||x||^2}$ then with B_λ as above

$$|\mathrm{Bias}_{\boldsymbol{\theta}}(\mathrm{SURE}(f, X))| \leq 2\, B_\lambda.$$

If the order conditions for the shrinkage bound hold and if $\lambda \in [0, 2(\mathrm{Tr}(\Sigma) - 2\kappa)]$ then this bound is of order $o(d)$.

Let $\mathrm{Y} \in \mathbb{R}^d$ be mean zero with positive definite covariance matrix Σ having entries $\sigma_{ij} = \mathrm{Cov}(Y_i, Y_j)$, we say the collection of vectors $\{\mathrm{Y}^{ij} : i, j \text{ such that } \sigma_{ij} \neq 0\}$ in $\mathbb{R}^d$ has the multivariate Y-zero bias distribution when

$$\mathbb{E}[\langle \mathrm{Y}, \mathrm{f}(\mathrm{Y})\rangle] = \mathbb{E}\left[\sum_{i,j=1}^{d} \sigma_{ij}\partial_j f_i(\mathrm{Y}^{ij})\right] =: \mathbb{E}[\langle \Sigma, \nabla \mathrm{f}(\mathrm{Y}^*)\rangle]$$

for all f in suitable Sobolev space.

Advantage: no continuity of the distribution of Y required.

This is a variant of the definition in *Goldstein, R. 2005*.

Zero bias consequence: shrinkage

Let $X = Y + \boldsymbol{\theta}$ where $Y \in \mathbb{R}^d$ has covariance matrix Σ with largest eigenvalue κ, and suppose that for all pairs i, j such that $\sigma_{ij} \neq 0$ the zero bias vectors X^{ij} exist. Then with $f(x) = \frac{x}{||x||^2}$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}||S_\lambda(X) - \boldsymbol{\theta}||^2 &= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda \langle X - \boldsymbol{\theta}, f(X)\rangle + \frac{\lambda^2}{||X||^2}\right\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda \langle \Sigma, \nabla f(X^*)\rangle + \frac{\lambda^2}{||X||^2}\right\} \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left\{||X - \boldsymbol{\theta}||^2 - 2\lambda \langle \Sigma, \nabla f(X)\rangle + \frac{\lambda^2}{||X||^2} + 2R\right\}
\end{aligned}$$

with $R = \mathbb{E}_{\boldsymbol{\theta}}\left\{\sum_{i,j=1}^d \sigma_{ij}[\partial_j f_i(X) - \partial_j f_i(X^{ij})]\right\}$.

We have already seen that $\mathbb{E}_{\boldsymbol{\theta}}[\langle \Sigma, \nabla \mathrm{f}(\mathrm{X})\rangle] \geq \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\mathrm{Tr}(\Sigma)-2\kappa}{||\mathrm{X}||^2}\right]$. We set

$$B_\lambda^* = \left|\mathbb{E}_{\boldsymbol{\theta}}\left\{\sum_{i,j=1}^{d} \sigma_{ij}[\partial_j f_i(\mathrm{X}) - \partial_j f_i(\mathrm{X}^{ij})]\right\}\right|.$$

Then it follows that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}||S_\lambda(\mathrm{X}) - \boldsymbol{\theta}||^2 \quad &\leq \quad \mathbb{E}_{\boldsymbol{\theta}}||\mathrm{X} - \boldsymbol{\theta}||^2 + \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\lambda}{||\mathrm{X}||^2}\left(\lambda - 2\left(\mathrm{Tr}(\Sigma) - 2\kappa\right)\right)\right] \\ &\quad +2B_\lambda^*.\end{aligned}$$

Example: Student distribution

$X = Y + \theta$ in $\mathbb{R}^d$ with Y from the family of multivariate central Student-t distributions, with $k \geq 5$ degrees of freedom, shape given by the identity matrix in $\mathbb{R}^{d\times d}$ and $d = 2m \geq 6$, even.

The covariance matrix is then $\Sigma = \sigma^2 \,\mathrm{Id}$.

We can write Y as

$$Y_\gamma = \gamma^{-1/2}\sigma N$$

with

$$N \sim \mathcal{N}_d(0, \mathrm{Id})$$

mixed over

$$\gamma \sim \Gamma(k/2, k/2).$$

Zero bias coupling for the Student

Write Y as

$$Y_\gamma = \gamma^{-1/2}\sigma N$$

with $N \sim \mathcal{N}_d(0, \mathrm{Id})$ mixed over $\gamma \sim \Gamma(k/2, k/2)$.

For $i = 1, \ldots, d$, the zero bias vectors Y^i are given by the mixture Y_δ where $\delta \sim \Gamma(k/2 - 1, k/2)$.

Let $\epsilon \sim \Gamma(1, k/2)$ be independent of δ and N be independent of both, and set $\gamma = \delta + \epsilon$.

Then a zero bias coupling is achieved by

$$X = \theta + \frac{\sigma}{\sqrt{\delta + \epsilon}} N \quad \text{and} \quad X^i = \theta + \frac{\sigma}{\sqrt{\delta}} N, \qquad i = 1, \ldots, d.$$

With this coupling, if $\theta = 0$,

$$B_\lambda^* \leq \frac{2\lambda}{k}.$$

This bound is $o(d)$ when $\lambda = O(d)$ and $1/k = o(1)$.

If $\boldsymbol{\theta} \neq 0$,

$$B_\lambda^* \leq \frac{8\lambda(d+k-2)}{(d-2)k}.$$

This bound is $o(d)$ when $\lambda = O(d)$ and $1/k = o(1)$.

Zero bias consequence: SURE

We introduce

$$\mathrm{SURE}_z(\mathrm{f}, \mathrm{X}) := \mathrm{Tr}(\Sigma) + \|\mathrm{f}(\mathrm{X})\|^2 + 2\sum_{i,j=1}^{d} \sigma_{ij}\partial_j f_i(\mathrm{X}^{ij}).$$

This is another unbiased risk estimate.

The zero bias construction usually depends on $\boldsymbol{\theta}$, here unknown. Let $\mathrm{X} = \boldsymbol{\theta} + \mathrm{Y}$ where Y has mean zero, covariance Σ, and whose zero bias vectors exist. Then,

$$|\mathrm{Bias}_{\boldsymbol{\theta}}(\mathrm{SURE}(\mathrm{f}, \mathrm{X}))| \leq 2\left|\sum_{i,j=1}^{d} \sigma_{ij} E_\theta\left(\partial_j f_i(\mathrm{X}^{ij}) - \partial_j f_i(\mathrm{X})\right)\right|.$$

Corollary

If $X = Y + \boldsymbol{\theta}$, if zero bias vectors of Y exist, and if $f(x) = -\lambda \frac{x}{||x||^2}$ then with B^*_λ as above

$$|\mathrm{Bias}_{\boldsymbol{\theta}}(\mathrm{SURE}(f, X))| \leq 2\, B^*_\lambda.$$

Outline

1. Background: $\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(X)\rangle]$
   - Stein's method
   - Stein's shrinkage estimator
   - Stein's Unbiased Risk Estimate (SURE)

2. Non-Gaussian models
   - Paths to extension
   - General Stein kernels $\mathbb{E}[\langle X-\boldsymbol{\theta}, f(X)\rangle] = \mathbb{E}[\langle \mathcal{T}_{X-\boldsymbol{\theta}}, \nabla f(X)\rangle]$
     - Stein kernel consequence: shrinkage
     - Stein kernel consequence: SURE
   - Zero-biasing $\mathbb{E}[\langle Y, f(Y)\rangle] = \mathbb{E}[\langle \Sigma, \nabla f(Y^*)\rangle]$
     - Zero bias consequence: shrinkage
     - Zero bias consequence: SURE

3. What else

We also have ...

- looked at other examples, such as strongly log-concave random vectors
- looked at soft thresholding, $f_\lambda(\mathrm{x}) = S_\lambda(\mathrm{x}) - \mathrm{x}$ with

$$S_\lambda(\mathrm{x})_i = sgn(x_i)(|x_i| - \lambda)_+$$

giving for $\Sigma = \sigma^2 \mathrm{Id}$

$$\mathrm{SURE}\,(\mathrm{f}_\lambda, \mathrm{X}) = d\sigma^2 + \sum_{i=1}^{d} \min\{X_i^2, \lambda^2\} - 2 \cdot \mathrm{Card}\{i : |X_i| \leqslant \lambda\}$$

instead of $\mathrm{SURE}(\mathrm{f}, \mathrm{X}) = \mathrm{Tr}(\Sigma) + \|\mathrm{f}(\mathrm{X})\|^2 + 2\langle \Sigma, \nabla \mathrm{f}(\mathrm{X})\rangle$

- looked at stability results over sets of distributions.

We have not ...

- looked at covariance matrix estimation
- looked at more involved applications
- looked at other Stein characterisations
- ...