Foundations of machine learning

Shrinkage in the Normal means model

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Outline

- Setup: the Normal means model

$$X \sim N(\theta, I_k)$$

  and the canonical estimation problem with loss $\|\widehat{\theta} - \theta\|^2$.

- The James-Stein (JS) shrinkage estimator.
- Three ways to arrive at the JS estimator (almost):
  1. Reverse regression of θ_i on X_i.
  2. Empirical Bayes: random effects model for θ_i.
  3. Shrinkage factor minimizing Stein's Unbiased Risk Estimate.
- Proof that JS uniformly dominates X as estimator of θ.
- Bonus slides: The Normal means model as asymptotic approximation.

Takeaways for this part of class

- Shrinkage estimators trade off variance and bias.
- In multi-dimensional problems, we can estimate the optimal degree of shrinkage.
- Three intuitions that lead to the JS-estimator:
  1. Predict θ_i given $X_i \Rightarrow$ reverse regression.
  2. Estimate distribution of the $\theta_i \Rightarrow$ empirical Bayes.
  3. Find shrinkage factor that minimizes estimated risk.
- Some calculus allows us to derive the risk of JS-shrinkage
  $\Rightarrow$ better than MLE, no matter what the true θ is.

The Normal means model
Setup

- $\theta \in \mathbb{R}^k$
- $\varepsilon \sim N(0, I_k)$
- $X = \theta + \varepsilon \sim N(\theta, I_k)$
- Estimator: $\widehat{\theta} = \widehat{\theta}(X)$
- Loss: squared error

$$L(\widehat{\theta}, \theta) = \sum_i (\widehat{\theta}_i - \theta_i)^2$$

- Risk: mean squared error

$$R(\widehat{\theta}, \theta) = E_\theta \left[L(\widehat{\theta}, \theta) \right] = \sum_i E_\theta \left[(\widehat{\theta}_i - \theta_i)^2 \right].$$

Two estimators

- Canonical estimator: maximum likelihood,

$$\widehat{\theta}^{ML} = X$$

- Risk function

$$R(\widehat{\theta}^{ML}, \theta) = \sum_i E_\theta \left[\varepsilon_i^2\right] = k.$$

- James-Stein shrinkage estimator

$$\widehat{\theta}^{JS} = \left(1 - \frac{(k-2)/k}{\overline{X^2}}\right) \cdot X.$$

- Celebrated result: uniform risk dominance; for all θ

$$R(\widehat{\theta}^{JS}, \theta) < R(\widehat{\theta}^{ML}, \theta) = k.$$

First motivation of JS: Regression perspective

- We will discuss three ways to motivate the JS-estimator (up to degrees of freedom correction).
- Consider estimators of the form
$$\widehat{\theta}_i = c \cdot X_i$$
or
$$\widehat{\theta}_i = a + b \cdot X_i.$$
- How to choose c or (a,b)?
- Two particular possibilities:
  1. Maximum likelihood: $c = 1$
  2. James-Stein: $c = \left(1 - \frac{(k-2)/k}{\overline{X^2}}\right)$

Practice problem (Infeasible estimator)

- Suppose you knew $X_1, \ldots, X_k$ as well as $\theta_1, \ldots, \theta_k$,
- but are constrained to use an estimator of the form $\widehat{\theta}_i = c \cdot X_i$.

1. Find the value of c that minimizes loss.
2. For estimators of the form $\widehat{\theta}_i = a + b \cdot X_i$, find the values of a and b that minimize loss.

Solution

- First problem:

$$c^* = \operatorname*{argmin}_c \sum_i (c \cdot X_i - \theta_i)^2$$

- Least squares problem!
- First order condition:

$$0 = \sum_i (c^* \cdot X_i - \theta_i) \cdot X_i.$$

- Solution

$$c^* = \frac{\sum X_i \theta_i}{\sum_i X_i^2}.$$

Solution continued

- Second problem:

$$(a^*, b^*) = \underset{a,b}{\operatorname{argmin}} \sum_i (a + b \cdot X_i - \theta_i)^2$$

- Least squares problem again!
- First order conditions:

$$0 = \sum_i (a^* + b^* \cdot X_i - \theta_i)$$
$$0 = \sum_i (a^* + b^* \cdot X_i - \theta_i) \cdot X_i.$$

- Solution

$$b^* = \frac{\sum (X_i - \overline{X}) \cdot (\theta_i - \overline{\theta})}{\sum_i (X_i - \overline{X})^2} = \frac{s_{X\theta}}{s_X^2}, \quad a^* + b^* \cdot \overline{X} = \overline{\theta}$$

Regression and reverse regression

- Recall $X_i = \theta_i + \varepsilon_i$, $E[\varepsilon_i|\theta_i] = 0$, $\mathrm{Var}(\varepsilon_i) = 1$.
- **Regression** of X on θ: Slope

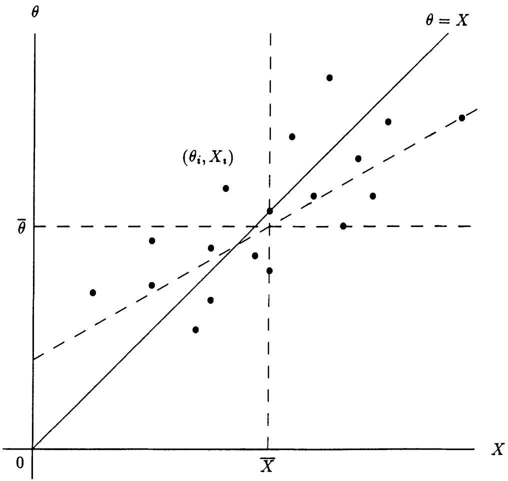
$$\frac{s_{X\theta}}{s_\theta^2} = 1 + \frac{s_{\varepsilon\theta}}{s_\theta^2} \approx 1.$$

- For optimal shrinkage, we want to predict θ given X, not the other way around!
- **Reverse regression** of θ on X: Slope

$$\frac{s_{X\theta}}{s_X^2} = \frac{s_\theta^2 + s_{\varepsilon\theta}}{s_\theta^2 + 2s_{\varepsilon\theta} + s_\varepsilon^2} \approx \frac{s_\theta^2}{s_\theta^2 + 1}.$$

- Interpretation: "signal to (signal plus noise) ratio" < 1.

Illustration



Expectations

Practice problem

1. Calculate the expectations of

$$\overline{X} = \frac{1}{k}\sum_i X_i, \quad \overline{X^2} = \frac{1}{k}\sum_i X_i^2,$$

and

$$s_X^2 = \frac{1}{k}\sum_i (X_i - \overline{X})^2 = \overline{X^2} - \overline{X}^2$$

2. Calculate the expected numerator and denominator of c^* and b^*.

Solution

- $E[\overline{X}] = \overline{\theta}$
- $E[\overline{X^2}] = \overline{\theta^2} + 1$
- $E[s_X^2] = \overline{\theta^2} - \overline{\theta}^2 + 1 = s_\theta^2 + 1$
- $c^* = (\overline{X\theta})/(\overline{X^2})$, and $E[\overline{X\theta}] = \overline{\theta^2}$. Thus

$$c^* \approx \frac{\overline{\theta^2}}{\overline{\theta^2} + 1}.$$

- $b^* = s_{X\theta}/s_X^2$, and $E[s_{X\theta}] = s_\theta^2$. Thus

$$b^* \approx \frac{s_\theta^2}{s_\theta^2 + 1}.$$

Feasible analog estimators

Practice problem

Propose feasible estimators of c^* and b^*.

A solution

- Recall:
  - $c^* = \frac{\overline{X\theta}}{\overline{X^2}}$
  - $\overline{\theta\varepsilon} \approx 0$, $\overline{\varepsilon^2} \approx 1$.
  - Since $X_i = \theta_i + \varepsilon_i$,

$$\overline{X\theta} = \overline{X^2} - \overline{X\varepsilon} = \overline{X^2} - \overline{\theta\varepsilon} - \overline{\varepsilon^2} \approx \overline{X^2} - 1$$

- Thus:

$$c^* = \frac{\overline{X^2} - \overline{\theta\varepsilon} - \overline{\varepsilon^2}}{\overline{X^2}} \approx \frac{\overline{X^2} - 1}{\overline{X^2}} = 1 - \frac{1}{\overline{X^2}} =: \widehat{c}.$$

Solution continued

- Similarly:
  - $b^* = \frac{s_{X\theta}}{s_X^2}$
  - $s_{\theta\varepsilon} \approx 0$, $s_\varepsilon^2 \approx 1$.
  - Since $X_i = \theta_i + \varepsilon_i$,
$$s_{X\theta} = s_X^2 - s_{X\varepsilon} = s_X^2 - s_{\theta\varepsilon} - s_\varepsilon^2 \approx s_X^2 - 1$$
- Thus:
$$b^* = \frac{s_X^2 - s_{\theta\varepsilon} - s_\varepsilon^2}{s_X^2} \approx \frac{s_X^2 - 1}{s_X^2} = 1 - \frac{1}{s_X^2} =: \widehat{b}$$

James-Stein shrinkage

- We have almost derived the James-Stein shrinkage estimator.
- Only difference: degree of freedom correction
- Optimal corrections:

$$c^{JS} = 1 - \frac{(k-2)/k}{\overline{X^2}},$$

and

$$b^{JS} = 1 - \frac{(k-3)/k}{s_X^2}.$$

- Note: if $\theta = 0$, then $\sum_i X_i^2 \sim \chi_k^2$.
- Then, by properties of inverse χ^2 distributions

$$E\left[\frac{1}{\sum_i X_i^2}\right] = \frac{1}{k-2},$$

so that $E\left[c^{JS}\right] = 0$.

Positive part JS-shrinkage

- The estimated shrinkage factors can be negative.
- $c^{JS} < 0$ iff
$$\sum_i X_i^2 < k - 2.$$
- Better estimator: restrict to $c \geq 0$.
- "Positive part James-Stein estimator:"
$$\widehat{\theta}^{JS+} = \max\left(0, 1 - \frac{(k-2)/k}{\overline{X^2}}\right) \cdot X.$$
- Dominates James-Stein.
- We will focus on the JS-estimator for analytical tractability.

Second motivation of JS: Parametric empirical Bayes
Setup

- As before: $\theta \in \mathbb{R}^k$
- $X|\theta \sim N(\theta, I_k)$
- Loss $L(\widehat{\theta}, \theta) = \sum_i (\widehat{\theta}_i - \theta_i)^2$
- Now add an additional conceptual layer:
  Think of θ_i as i.i.d. draws from some distribution.
- "Random effects vs. fixed effects"
- Let's consider $\theta_i \sim^{iid} N(0, \tau^2)$,
  where τ^2 is unknown.

Practice problem

- Derive the marginal distribution of X given τ^2.
- Find the maximum likelihood estimator of τ^2.
- Find the conditional expectation of θ given X and τ^2.
- Plug in the maximum likelihod estimator of τ^2 to get the empirical Bayes estimator of θ.

Solution

- Marginal distribution:

$$X \sim N\left(0, (\tau^2+1)\cdot I_k\right)$$

- Maximum likelihood estimator of τ^2:

$$\begin{aligned}\widehat{\tau^2} =& \underset{t^2}{\operatorname{argmax}} \; -\frac{1}{2}\sum_i \left(\log(\tau^2+1) + \frac{X_i^2}{(\tau^2+1)}\right) \\ =& \overline{X^2} - 1\end{aligned}$$

- Conditional expectation of θ_i given X_i, τ^2:

$$\widehat{\theta}_i = \frac{\operatorname{Cov}(\theta_i, X_i)}{\operatorname{Var}(X_i)} \cdot X_i = \frac{\tau^2}{\tau^2+1} \cdot X_i.$$

- Plugging in $\widehat{\tau^2}$:

$$\widehat{\theta}_i = \left(1 - \frac{1}{\overline{X^2}}\right) \cdot X_i.$$

General parametric empirical Bayes
Setup

- Data X,
  parameters θ,
  hyper-parameters η

- Likelihood
$$X|\theta,\eta \sim f_{X|\theta}$$

- Family of priors
$$\theta|\eta \sim f_{\theta|\eta}$$

- Limiting cases:
  - $\theta = \eta$: Frequentist setup.
  - η has only one possible value: Bayesian setup.

Empirical Bayes estimation

- Marginal likelihood

$$f_{X|\eta}(x|\eta) = \int f_{X|\theta}(x|\theta) f_{\theta|\eta}(\theta|\eta) d\theta.$$

Has simple form when family of priors is conjugate.

- Estimator for hyper-parameter η: marginal MLE

$$\widehat{\eta} = \underset{\eta}{\text{argmax}}\ f_{X|\eta}(x|\eta).$$

- Estimator for parameter θ: pseudo-posterior expectation

$$\widehat{\theta} = E[\theta|X = x, \eta = \widehat{\eta}].$$

Third motivation of JS: Stein's Unbiased Risk Estimate

- Stein's lemma (simplified version):
- Suppose $X \sim N(\theta, I_k)$.
- Suppose $g(\cdot) : \mathbb{R}^k \to \mathbb{R}$ is differentiable and $E[|g'(X)|] < \infty$.
- Then
$$E[(X - \theta) \cdot g(X)] = E[\nabla g(X)].$$
- Note:
    - θ shows up in the expression on the LHS, but not on the RHS
    - Unbiased estimator of the RHS: $\nabla g(X)$

Practice problem

Prove this.

Hints:

1. Show that the standard Normal density $\varphi(\cdot)$ satisfies

$$\varphi'(x) = -x \cdot \varphi(x).$$

2. Consider each component i separately and use integration by parts.

Solution

- Recall that $\varphi(x) = (2\pi)^{-0.5} \cdot \exp(-x^2/2)$.
  Differentiation immediately yields the first claim.

- Consider the component $i = 1$; the others follow similarly. Then

$$
\begin{aligned}
& E[\partial_{x_1} g(X)] = \\
= & \int_{x_2,\ldots x_k} \int_{x_1} \partial_{x_1} g(x_1,\ldots,x_k) \cdot \varphi(x_1-\theta_1) \cdot \prod_{i=2}^{k} \varphi(x_i-\theta_i) dx_1 \ldots dx_k \\
= & \int_{x_2,\ldots x_k} \int_{x_1} g(x_1,\ldots,x_k) \cdot (-\partial_{x_1} \varphi(x_1-\theta_1)) \cdot \prod_{i=2}^{k} \varphi(x_i-\theta_i) dx_1 \ldots dx_k \\
= & \int_{x_2,\ldots x_k} \int_{x_1} g(x_1,\ldots,x_k) \cdot (x_1-\theta_1)\varphi(x_1-\theta_1) \cdot \prod_{i=2}^{k} \varphi(x_i-\theta_i) dx_1 \ldots dx_k \\
= & E[(X_1-\theta_1) \cdot g(X)].
\end{aligned}
$$

- Collecting the components $i = 1, \ldots, k$ yields

$$E[(X-\theta) \cdot g(X)] = E[\nabla g(X)].$$

Stein's representation of risk

- Consider a general estimator for θ of the form $\widehat{\theta} = \widehat{\theta}(X) = X + g(X)$, for differentiable g.
- Recall that the risk function is defined as

$$R(\widehat{\theta}, \theta) = \sum_i E[(\widehat{\theta}_i - \theta_i)^2].$$

- We will show that this risk function can be rewritten as

$$R(\widehat{\theta}, \theta) = k + \sum_i \left(E[g_i(X)^2] + 2E[\partial_{x_i} g_i(X)] \right).$$

Practice problem

- Interpret this expression.
- Propose an unbiased estimator of risk, based on this expression.

Answer

- The expression of risk has 3 components:
  1. k is the risk of the canonical estimator $\widehat{\theta} = X$, corresponding to $g \equiv 0$.
  2. $\sum_i E[g_i(X)^2] = \sum_i E[(\widehat{\theta}_i - X_i)^2]$ is the sample sum of squared errors.
  3. $\sum_i E[\partial_{x_i} g_i(X)]$ can be thought of as a penalty for overfitting.
- We thus can think of this expression as giving a "penalized least squares" objective.
- The sample analog expression gives "Stein's Unbiased Risk Estimate" (SURE)

$$\widehat{R} = k + \sum_i \left(\widehat{\theta}_i - X_i\right)^2 + 2 \cdot \sum_i \partial_{x_i} g_i(X).$$

- We will use Stein's representation of risk in 2 ways:
    1. To derive feasible optimal shrinkage parameter using its sample analog (SURE).
    2. To prove uniform dominance of JS using population version.

Practice problem

Prove Stein's representation of risk.
Hints:

- Add and subtract X_i in the expression defining $R(\widehat{\theta}, \theta)$.
- Use Stein's lemma.

Solution

$$
\begin{aligned}
R(\theta) &= \sum_i E\left[(\widehat{\theta}_i - X_i + X_i - \theta_i)^2\right] \\
&= \sum_i E\left[(X_i - \theta_i)^2 \qquad + (\widehat{\theta}_i - X_i)^2 \qquad + 2(\widehat{\theta}_i - X_i)\cdot(X_i - \theta_i)\right] \\
&= \sum_i 1 \qquad + E\left[g_i(X)^2\right] \qquad + 2E\left[g_i(X)\cdot(X_i - \theta_i)\right] \\
&= \sum_i 1 \qquad + E\left[g_i(X)^2\right] \qquad + 2E\left[\partial_{x_i} g_i(X)\right],
\end{aligned}
$$

where Stein's lemma was used in the last step.

Using SURE to pick the tuning parameter

- First use of SURE: To pick tuning parameters, as an alternative to cross-validation or marginal likelihood maximization.
- Simple example: Linear shrinkage estimation

$$\widehat{\theta} = c \cdot X.$$

Practice problem

- Calculate Stein's unbiased risk estimate for $\widehat{\theta}$.
- Find the coefficient c minimizing estimated risk.

Solution

- When $\widehat{\theta} = c \cdot X$,
  then $g(X) = \widehat{\theta} - X = (c-1) \cdot X$,
  and $\partial_{x_i} g_i(X) = c - 1$.

- Estimated risk:
  $$\widehat{R} = k + (1-c)^2 \cdot \sum_i X_i^2 + 2k \cdot (c-1).$$

- First order condition for minimizing $\widehat{R}$:
  $$k = (1-c^*) \cdot \sum_i X_i^2.$$

- Thus
  $$c^* = 1 - \frac{1}{\overline{X^2}}.$$

- Once again: Almost the JS estimator, up to degrees of freedom correction!

Celebrated result: Dominance of the JS-estimator

- We next use the population version of SURE to prove uniform dominance of the JS-estimator relative to maximum likelihood.

- Recall that the James-Stein estimator was defined as

$$\widehat{\theta}^{JS} = \left(1 - \frac{(k-2)/k}{\overline{X^2}}\right) \cdot X.$$

- Claim: The JS-estimator has uniformly lower risk than $\widehat{\theta}^{ML} = X$.

Practice problem

Prove this, using Stein's representation of risk.

Solution

- The risk of $\widehat{\theta}^{ML}$ is equal to k.

- For JS, we have

$$g_i(X) = \widehat{\theta}_i^{JS} - X_i = -\frac{k-2}{\sum_j X_j^2} \cdot X_i, \quad \text{and}$$

$$\partial_{x_i} g_i(X) = \frac{k-2}{\sum_j X_j^2} \cdot \left(-1 + \frac{2X_i^2}{\sum_j X_j^2} \right).$$

- Summing over components gives

$$\sum_i g_i(X)^2 = \frac{(k-2)^2}{\sum_j X_j^2}, \quad \text{and}$$

$$\sum_i \partial_{x_i} g_i(X) = -\frac{(k-2)^2}{\sum_j X_j^2}.$$

Solution continued

- Plugging into Stein's expression for risk then gives

$$
\begin{aligned}
R(\widehat{\theta}^{JS}, \theta) =& k + E\left[\sum_i g_i(X)^2 + 2\sum_i \partial_{x_i} g_i(X)\right] \\
=& k + E\left[\frac{(k-2)^2}{\sum_i X_i^2} - 2\frac{(k-2)^2}{\sum_j X_j^2}\right] \\
=& \, k - E\left[\frac{(k-2)^2}{\sum_i X_i^2}\right].
\end{aligned}
$$

- The term $\frac{(k-2)^2}{\sum_i X_i^2}$ is always positive (for $k \geq 3$), and thus so is its expectation. Uniform dominance immediately follows.

- Pretty cool, no?

References

- Textbook introduction:
Wasserman, L. (2006). All of nonparametric statistics. *Springer Science & Business Media, chapter 7.*

- Reverse regression perspective:
Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. Statistical Science, *pages 147–155.*

- Parametric empirical Bayes:
  Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. Journal of the American Statistical Association, *78(381):pp. 47–55.*

  Lehmann, E. L. and Casella, G. (1998). Theory of point estimation, *volume 31. Springer, section 4.6.*

- Stein's Unbiased Risk Estimate:
  Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. The Annals of Statistics, *9(6):1135–1151.*

  Lehmann, E. L. and Casella, G. (1998). Theory of point estimation, *volume 31. Springer, sections 5.2, 5.4, 5.5.*