High Dimensional Predictive Inference

Ed George
University of Pennsylvania
(joint work with Larry Brown, Feng Liang, and Xinyi Xu)

Conference on Predictive Inference and Its Applications
Iowa State University
May 7, 2018

I. The Hunt for Shrinkage Estimators Begins

I. The Hunt for Shrinkage Estimators Begins

- Canonical Problem: Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2$$

I. The Hunt for Shrinkage Estimators Begins

- Canonical Problem: Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2$$

- $\hat{\mu}_{MLE}(X) = X$ is MLE, best invariant and minimax with constant risk $R_Q(\mu, \hat{\mu}_{MLE}) \equiv p$.

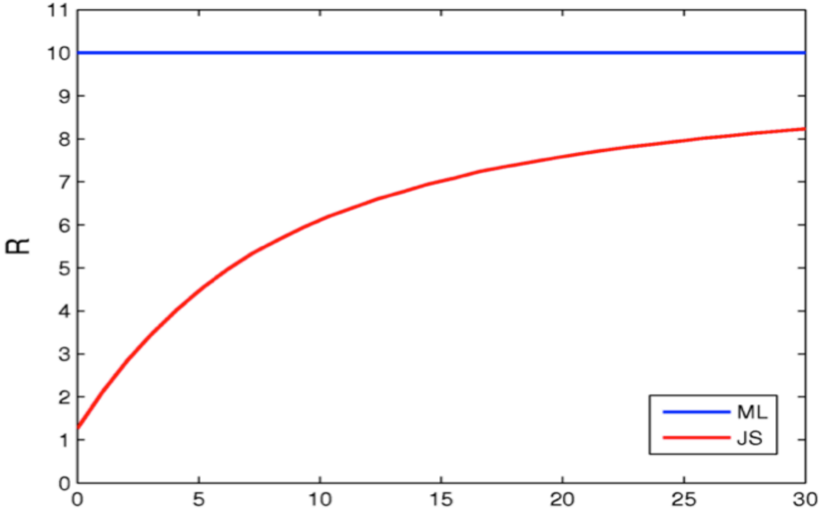
I. The Hunt for Shrinkage Estimators Begins

- Canonical Problem: Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under
$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2$$
- $\hat{\mu}_{MLE}(X) = X$ is MLE, best invariant and minimax with constant risk $R_Q(\mu, \hat{\mu}_{MLE}) \equiv p$.
- A Shocking Discovery: $\hat{\mu}_{MLE}$ is inadmissible when $p \geq 3$. (Stein 1956)

I. The Hunt for Shrinkage Estimators Begins

- Canonical Problem: Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under
$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2$$
- $\hat{\mu}_{MLE}(X) = X$ is MLE, best invariant and minimax with constant risk $R_Q(\mu, \hat{\mu}_{MLE}) \equiv p$.
- A Shocking Discovery: $\hat{\mu}_{MLE}$ is inadmissible when $p \geq 3$. (Stein 1956)
- An Explicit Better Estimator Appears: The James-Stein estimator
$$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X$$
(James and Stein 1961)

- The risk of $\hat{\mu}_{MLE}$ and the risk of $\hat{\mu}_{JS}$ various values of μ



$\|\mu\|^2$

- Stein (1962) suggests an empirical Bayes motivation for $\hat{\mu}_{JS}$.
The focus of the hunt turns to Bayes.

- Stein (1962) suggests an empirical Bayes motivation for $\hat{\mu}_{JS}$. The focus of the hunt turns to Bayes.

- For a prior $\pi(\mu)$, the Bayes rule under $R_Q(\mu, \hat{\mu})$ is

$$\hat{\mu}_\pi(X) = E_\pi(\mu \mid X)$$

- Stein (1962) suggests an empirical Bayes motivation for $\hat{\mu}_{JS}$. The focus of the hunt turns to Bayes.

- For a prior $\pi(\mu)$, the Bayes rule under $R_Q(\mu, \hat{\mu})$ is

$$\hat{\mu}_\pi(X) = E_\pi(\mu \mid X)$$

- Remark: The (formal) Bayes rule under $\pi_U(\mu) \equiv 1$ is

$$\hat{\mu}_U(X) \equiv \hat{\mu}_{MLE}(X) = X$$

- $\hat{\mu}_H(X)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $\hat{\mu}_U$ when $p \geq 3$. (Stein 1974)

- $\hat{\mu}_H(X)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

  dominates $\hat{\mu}_U$ when $p \geq 3$. (Stein 1974)

- $\hat{\mu}_a(X)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu \mid s \sim N_p\left(0, s\, I\right), \quad s \sim (1+s)^{a-2}$$

  dominates $\hat{\mu}_U$ and is proper Bayes when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Strawderman 1971)

- $\hat{\mu}_H(X)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $\hat{\mu}_U$ when $p \geq 3$. (Stein 1974)

- $\hat{\mu}_a(X)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu \mid s \sim N_p\left(0, s\, I\right), \quad s \sim (1+s)^{a-2}$$

dominates $\hat{\mu}_U$ and is proper Bayes when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Strawderman 1971)

- A Unifying Phenomenon: These domination results can be attributed to properties of the marginal distribution of X under π_H and π_a.

- The Bayes rule under $\pi(\mu)$ can be expressed as

$$\hat{\mu}_\pi(X) = E_\pi(\mu \mid X) = X + \nabla \log m_\pi(X)$$

where

$$m_\pi(X) \propto \int e^{-(X-\mu)^2/2}\, \pi(\mu)\, d\mu$$

is the marginal of X under $\pi(\mu)$. ($\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p})'$)
(Brown 1971)

- The Bayes rule under $\pi(\mu)$ can be expressed as

$$\hat{\mu}_\pi(X) = E_\pi(\mu \mid X) = X + \nabla \log m_\pi(X)$$

where

$$m_\pi(X) \propto \int e^{-(X-\mu)^2/2}\, \pi(\mu)\, d\mu$$

is the marginal of X under $\pi(\mu)$. $(\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p})')$
(Brown 1971)

- The risk improvement of $\hat{\mu}_\pi(X)$ over $\hat{\mu}_U(X)$ can be expressed as

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) = E_\mu \left[\|\nabla \log m_\pi(X)\|^2 - 2\frac{\nabla^2 m_\pi(X)}{m_\pi(X)} \right]$$

$$= E_\mu \left[-4\nabla^2 \sqrt{m_\pi(X)} / \sqrt{m_\pi(X)} \right]$$

$(\nabla^2 = \sum_i \frac{\partial^2}{\partial x_i^2})$ (Stein 1974, 1981)

- That $\hat{\mu}_H(X)$ dominates $\hat{\mu}_U$ when $p \geq 3$, follows from the fact that the marginal $m_\pi(X)$ under π_H is superharmonic, i.e.

$$\nabla^2 m_\pi(X) \leq 0$$

- That $\hat{\mu}_H(X)$ dominates $\hat{\mu}_U$ when $p \geq 3$, follows from the fact that the marginal $m_\pi(X)$ under π_H is superharmonic, i.e.

$$\nabla^2 m_\pi(X) \leq 0$$

- That $\hat{\mu}_a(X)$ dominates $\hat{\mu}_U$ when $p \geq 5$ (and conditions on a), follows from the fact that the sqrt of the marginal under π_a is superharmonic, i.e.

$$\nabla^2 \sqrt{m_\pi(X)} \leq 0$$

(Fourdrinier, Strawderman and Wells 1998)

II. The Prediction Problem

II. The Prediction Problem

- Observe $X \mid \mu \sim N_p(\mu, v_x I)$ and predict $Y \mid \mu \sim N_p(\mu, v_y I)$
  - Given μ, Y is independent of X
  - v_x and v_y are known (for now)

II. The Prediction Problem

- Observe $X \mid \mu \sim N_p(\mu, v_x I)$ and predict $Y \mid \mu \sim N_p(\mu, v_y I)$
  - Given μ, Y is independent of X
  - v_x and v_y are known (for now)

- The Problem: To estimate $p(y \mid \mu)$ by $q(y \mid x)$.

II. The Prediction Problem

- Observe $X \mid \mu \sim N_p(\mu, v_x I)$ and predict $Y \mid \mu \sim N_p(\mu, v_y I)$
  - Given μ, Y is independent of X
  - v_x and v_y are known (for now)

- The Problem: To estimate $p(y \mid \mu)$ by $q(y \mid x)$.

- Measure closeness by Kullback-Leibler loss,

$$L(\mu, q(y \mid x)) = \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{q(y \mid x)} dy$$

II. The Prediction Problem

- Observe $X \mid \mu \sim N_p(\mu, v_x I)$ and predict $Y \mid \mu \sim N_p(\mu, v_y I)$
  - Given μ, Y is independent of X
  - v_x and v_y are known (for now)

- The Problem: To estimate $p(y \mid \mu)$ by $q(y \mid x)$.

- Measure closeness by Kullback-Leibler loss,

$$L(\mu, q(y \mid x)) = \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{q(y \mid x)} dy$$

- Risk function

$$R_{KL}(\mu, q) = \int L(\mu, q(y \mid x))\, p(x \mid \mu)\, dx = E_\mu[L(\mu, q(y \mid X))]$$

Bayes Rules for the Prediction Problem

Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule under $R_{KL}(\mu, q)$ is

$$p_{\pi}(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_{\pi}[p(y \mid \mu)|X]$$

Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule under $R_{KL}(\mu, q)$ is

$$p_\pi(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_\pi[p(y \mid \mu)|X]$$

- Let $p_U(y \mid x)$ denote the (formal) Bayes rule under $\pi_U(\mu) \equiv 1$.

Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule under $R_{KL}(\mu, q)$ is

$$p_\pi(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_\pi[p(y \mid \mu)|X]$$

- Let $p_U(y \mid x)$ denote the (formal) Bayes rule under $\pi_U(\mu) \equiv 1$.
- $p_U(y \mid x)$ dominates $p(y \mid \hat{\mu} = x)$, the naive "plug-in" predictive distribution. (Aitchison 1975)

Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule under $R_{KL}(\mu, q)$ is

$$p_\pi(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_\pi[p(y \mid \mu)|X]$$

- Let $p_U(y \mid x)$ denote the (formal) Bayes rule under $\pi_U(\mu) \equiv 1$.
- $p_U(y \mid x)$ dominates $p(y \mid \hat{\mu} = x)$, the naive "plug-in" predictive distribution. (Aitchison 1975)
- $p_U(y \mid x)$ is best invariant and minimax with constant risk. (Murray 1977, Ng 1980, Barron and Liang 2003)

Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule under $R_{KL}(\mu, q)$ is

$$p_\pi(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_\pi[p(y \mid \mu)|X]$$

- Let $p_U(y \mid x)$ denote the (formal) Bayes rule under $\pi_U(\mu) \equiv 1$.
- $p_U(y \mid x)$ dominates $p(y \mid \hat{\mu} = x)$, the naive "plug-in" predictive distribution. (Aitchison 1975)
- $p_U(y \mid x)$ is best invariant and minimax with constant risk. (Murray 1977, Ng 1980, Barron and Liang 2003)
- Shocking Fact: $p_U(y \mid x)$ is inadmissible when $p \geq 3$.

- $p_H(y \mid x)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $p_U(y \mid x)$ when $p \geq 3$. (Komaki 2001)

- $p_H(y \mid x)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $p_U(y \mid x)$ when $p \geq 3$. (Komaki 2001)

- $p_a(y \mid x)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu \mid s \sim N_p\left(0, s\, v_0 I\right), \quad s \sim (1+s)^{a-2},$$

dominates $p_U(y \mid x)$ and is proper Bayes when $v_x \leq v_0$ and when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Liang 2002)

- $p_H(y \mid x)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $p_U(y \mid x)$ when $p \geq 3$. (Komaki 2001)

- $p_a(y \mid x)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu \mid s \sim N_p\left(0, s\, v_0 I\right), \quad s \sim (1+s)^{a-2},$$

dominates $p_U(y \mid x)$ and is proper Bayes when $v_x \leq v_0$ and when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Liang 2002)

- A Key Question: Are these domination results attributable to the properties of m_π?

A Key Representation for $p_\pi(y \mid x)$

A Key Representation for $p_\pi(y \mid x)$

- Let $m_\pi(x; v_x)$ denote the marginal of $X \mid \mu \sim N_p(\mu, v_x I)$ under $\pi(\mu)$.

A Key Representation for $p_\pi(y \mid x)$

- Let $m_\pi(x; v_x)$ denote the marginal of $X \mid \mu \sim N_p(\mu, v_x I)$ under $\pi(\mu)$.

- **Lemma**: The Bayes rule $p_\pi(y \mid x)$ can be expressed as

$$p_\pi(y \mid x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)}\, p_U(y \mid x)$$

  where

$$W = \frac{v_y X + v_x Y}{v_x + v_y} \sim N_p(\mu, v_w I)$$

A Key Representation for $p_\pi(y \mid x)$

- Let $m_\pi(x; v_x)$ denote the marginal of $X \mid \mu \sim N_p(\mu, v_x I)$ under $\pi(\mu)$.

- **Lemma**: The Bayes rule $p_\pi(y \mid x)$ can be expressed as

$$p_\pi(y \mid x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} \, p_U(y \mid x)$$

where

$$W = \frac{v_y X + v_x Y}{v_x + v_y} \sim N_p(\mu, v_w I)$$

- Using this, the risk improvement can be expressed as

$$R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) = \int \int p_{v_x}(x|\mu)\, p_{v_y}(y|\mu) \log \frac{p_\pi(y \mid x)}{p_U(y \mid x)} dx dy$$

$$= E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x)$$

An Analogue of Stein's Unbiased Estimate of Risk

An Analogue of Stein's Unbiased Estimate of Risk

- **Theorem**:

$$\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z;v) = E_{\mu,v}\left[\frac{\nabla^2 m_\pi(Z;v)}{m_\pi(Z;v)} - \frac{1}{2}\|\nabla \log m_\pi(Z;v)\|^2\right]$$

$$= E_{\mu,v}\left[2\nabla^2\sqrt{m_\pi(Z;v)}/\sqrt{m_\pi(Z;v)}\right]$$

An Analogue of Stein's Unbiased Estimate of Risk

- **Theorem**:

$$\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z;v) = E_{\mu,v}\left[\frac{\nabla^2 m_\pi(Z;v)}{m_\pi(Z;v)} - \frac{1}{2}\|\nabla \log m_\pi(Z;v)\|^2\right]$$
$$= E_{\mu,v}\left[2\nabla^2\sqrt{m_\pi(Z;v)}/\sqrt{m_\pi(Z;v)}\right]$$

- Proof relies on using the heat equation

$$\frac{\partial}{\partial v} m_\pi(z;v) = \frac{1}{2}\nabla^2 m_\pi(z;v),$$

Brown's representation and Stein's Lemma.

General Conditions for Minimax Prediction

General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.

General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.

- **Theorem**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if either of the following hold:
  1. $m_\pi(z; v)$ is superharmonic
  2. $\sqrt{m_\pi(z; v)}$ is superharmonic

General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.

- **Theorem**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if either of the following hold:
    1. $m_\pi(z; v)$ is superharmonic
    2. $\sqrt{m_\pi(z; v)}$ is superharmonic

- **Corollary**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if $\pi(\mu)$ is superharmonic.

General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.

- **Theorem**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if either of the following hold:
  1. $m_\pi(z; v)$ is superharmonic
  2. $\sqrt{m_\pi(z; v)}$ is superharmonic

- **Corollary**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if $\pi(\mu)$ is superharmonic.

- $p_\pi(y \mid x)$ will dominate $p_U(y \mid x)$ in the above results if the superharmonicity is strict on some interval.

Consequences of the General Minimax Conditions

Consequences of the General Minimax Conditions

- Because π_H is superharmonic, it is immediate that $p_H(y \mid x)$ dominates $p_U(y \mid x)$ and is minimax.

Consequences of the General Minimax Conditions

- Because π_H is superharmonic, it is immediate that $p_H(y \mid x)$ dominates $p_U(y \mid x)$ and is minimax.

- Because $\sqrt{m_a}$ is superharmonic (under suitable conditions on a), it is immediate that $p_a(y \mid x)$ dominates $p_U(y \mid x)$ and is minimax.

Consequences of the General Minimax Conditions

- Because π_H is superharmonic, it is immediate that $p_H(y \mid x)$ dominates $p_U(y \mid x)$ and is minimax.

- Because $\sqrt{m_a}$ is superharmonic (under suitable conditions on a), it is immediate that $p_a(y \mid x)$ dominates $p_U(y \mid x)$ and is minimax.

- It also follows that any of the improper superharmonic t-priors of Faith (1978) or any of the proper generalized t-priors of Fourdrinier, Strawderman and Wells (1998) yield Bayes rules that dominate $p_U(y \mid x)$ and are minimax.

III. Predictive "Shrinkage"

III. Predictive "Shrinkage"

- Our Lemma representation

$$p_H(y \mid x) = \frac{m_H(w; v_w)}{m_H(x; v_x)} \, p_U(y \mid x)$$

shows how $p_H(y \mid x)$ "shrinks $p_U(y \mid x)$ towards 0" by an adaptive multiplicative factor.

III. Predictive "Shrinkage"

- Our Lemma representation

$$p_H(y \mid x) = \frac{m_H(w; v_w)}{m_H(x; v_x)} \; p_U(y \mid x)$$

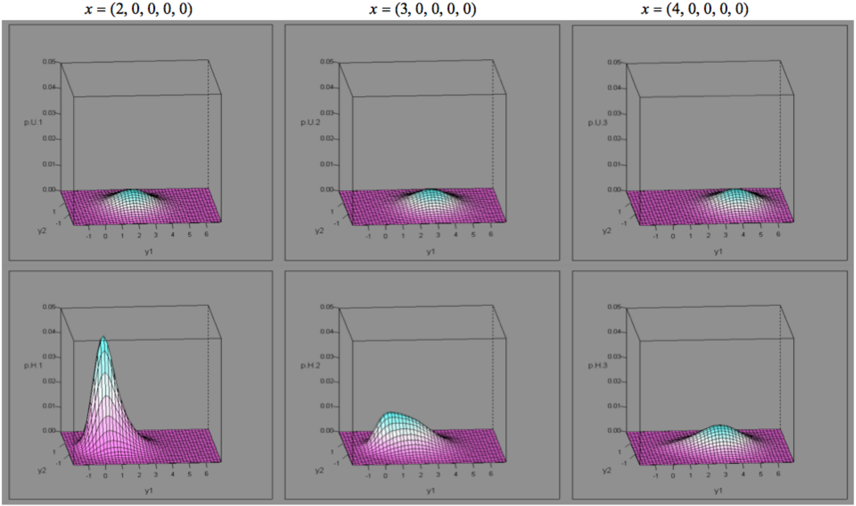
shows how $p_H(y \mid x)$ "shrinks $p_U(y \mid x)$ towards 0" by an adaptive multiplicative factor.

- Note the analogies with the Bayes rule $\hat{\mu}_\pi(X) = E_\pi(\mu \mid X)$ whose coordinates are

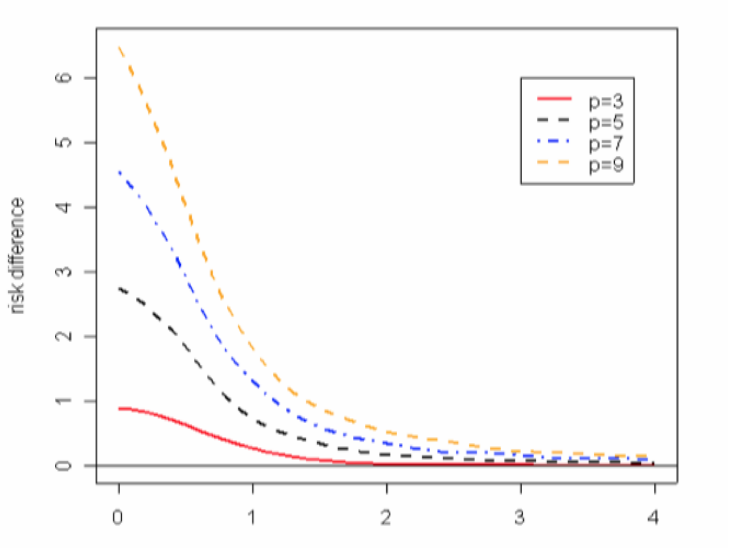
$$\left(1 + \frac{(\nabla \log m_\pi(X))_i}{X_i}\right) \; X_i$$

Predictive Shrinkage in Action

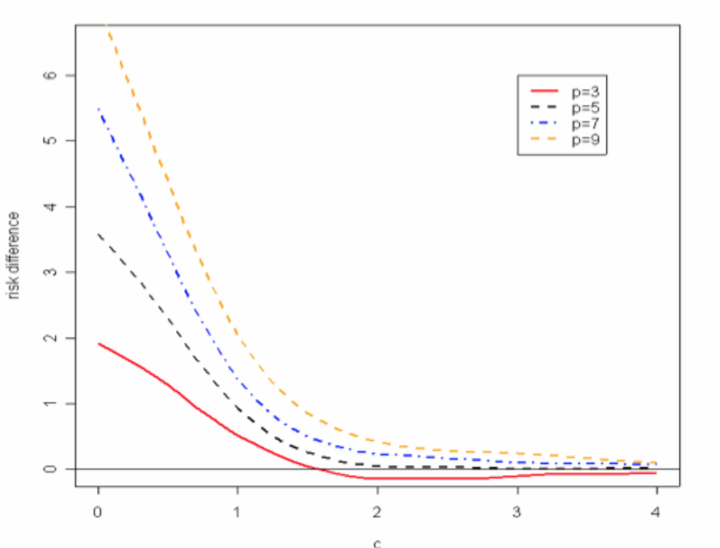
- The contrast between $p_U(y \mid x)$ and $p_H(y \mid x)$ for various values of x



- The risk function difference $[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_H)]$ is largest at $\mu = 0$, and then asymptotes to 0 as $\|\mu\| \to \infty$.



- The risk function difference $[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_a)]$ is largest at $\mu = 0$, and then asymptotes to 0 as $\|\mu\| \to \infty$.



Predictive Shrinkage Towards Points or Subspaces

Predictive Shrinkage Towards Points or Subspaces

- We can trivially modify the previous priors and predictive distributions to shrink towards an arbitrary point $b \in R^p$.

Predictive Shrinkage Towards Points or Subspaces

- We can trivially modify the previous priors and predictive distributions to shrink towards an arbitrary point $b \in R^p$.
- Consider the recentered prior

$$\pi^b(\mu) = \pi(\mu - b)$$

and corresponding recentered marginal

$$m_\pi^b(z; v) = m_\pi(z - b; v).$$

Predictive Shrinkage Towards Points or Subspaces

- We can trivially modify the previous priors and predictive distributions to shrink towards an arbitrary point $b \in R^p$.
- Consider the recentered prior

$$\pi^b(\mu) = \pi(\mu - b)$$

and corresponding recentered marginal

$$m_\pi^b(z; v) = m_\pi(z - b; v).$$

- This yields a predictive distribution

$$p_\pi^b(y \mid x) = \frac{m_\pi^b(w; v_w)}{m_\pi^b(x; v_x)} \, p_U(y \mid x)$$

that now shrinks $p_U(y \mid x)$ towards b rather than 0.

- More generally, we can shrink $p_U(y \mid x)$ towards any subspace B of R^p whenever π, and hence m_π, is spherically symmetric.

- More generally, we can shrink $p_U(y \mid x)$ towards any subspace B of R^p whenever π, and hence m_π, is spherically symmetric.
- Letting $P_B z$ be the projection of z onto B, shrinkage towards B is obtained by using the recentered prior

$$\pi^B(\mu) = \pi(\mu - P_B \mu)$$

which yields the reecentered marginal

$$m_\pi^B(z; v) := m_\pi(z - P_B z; v).$$

- More generally, we can shrink $p_U(y \mid x)$ towards any subspace B of R^p whenever π, and hence m_π, is spherically symmetric.
- Letting $P_B z$ be the projection of z onto B, shrinkage towards B is obtained by using the recentered prior

$$\pi^B(\mu) = \pi(\mu - P_B \mu)$$

which yields the reecentered marginal

$$m_\pi^B(z; v) := m_\pi(z - P_B z; v).$$

- This modification yields a predictive distribution

$$p_\pi^B(y \mid x) = \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)} \, p_U(y \mid x)$$

that now shrinks $p_U(y \mid x)$ towards B.

- More generally, we can shrink $p_U(y \mid x)$ towards any subspace B of R^p whenever π, and hence m_π, is spherically symmetric.
- Letting $P_B z$ be the projection of z onto B, shrinkage towards B is obtained by using the recentered prior

$$\pi^B(\mu) = \pi(\mu - P_B \mu)$$

which yields the reecentered marginal

$$m_\pi^B(z; v) := m_\pi(z - P_B z; v).$$

- This modification yields a predictive distribution

$$p_\pi^B(y \mid x) = \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)} \, p_U(y \mid x)$$

that now shrinks $p_U(y \mid x)$ towards B.
- If $m_\pi^B(z; v)$ satisfies any of our superharmonic conditions for minimaxity, then $p_\pi^B(y \mid x)$ will dominate $p_U(y \mid x)$ and be minimax.

Minimax Multiple Predictive Shrinkage

Minimax Multiple Predictive Shrinkage

- For any spherically symmetric prior, a set of subspaces $B_1, \ldots, B_N$, and corresponding probabilities $w_1, ..., w_N$, consider the recentered mixture prior

$$\pi_*(\mu) = \sum_{i=1}^{N} w_i \, \pi^{B_i}(\mu),$$

and corresponding recentered mixture marginal

$$m_*(z; v) = \sum_{1}^{N} w_i \, m_\pi^{B_i}(z; v).$$

Minimax Multiple Predictive Shrinkage

- For any spherically symmetric prior, a set of subspaces $B_1, \ldots, B_N$, and corresponding probabilities $w_1, ..., w_N$, consider the recentered mixture prior
$$\pi_*(\mu) = \sum_{i=1}^{N} w_i \, \pi^{B_i}(\mu),$$
and corresponding recentered mixture marginal
$$m_*(z; v) = \sum_{1}^{N} w_i \, m_\pi^{B_i}(z; v).$$
- Applying the $\hat{\mu}_\pi(X) = X + \nabla \log m_\pi(X)$ construction with $m_*(X; v)$ yields minimax multiple shrinkage estimators of μ. (George 1986)

- Applying the predictive construction with $m_*(z; v)$ yields

$$p_*(y \mid x) = \sum_{i=1}^{N} p(B_i \mid x)\, p_\pi^{B_i}(y \mid x)$$

where $p_\pi^{B_i}(y \mid x)$ is a single target predictive distribution and

$$p(B_i \mid x) = \frac{w_i m_\pi^{B_i}(x; v_x)}{\sum_{i=1}^{N} w_i m_\pi^{B_i}(x; v_x)}$$

is the posterior weight on the ith prior component.

- Applying the predictive construction with $m_*(z; v)$ yields

$$p_*(y \mid x) = \sum_{i=1}^{N} p(B_i \mid x)\, p_\pi^{B_i}(y \mid x)$$

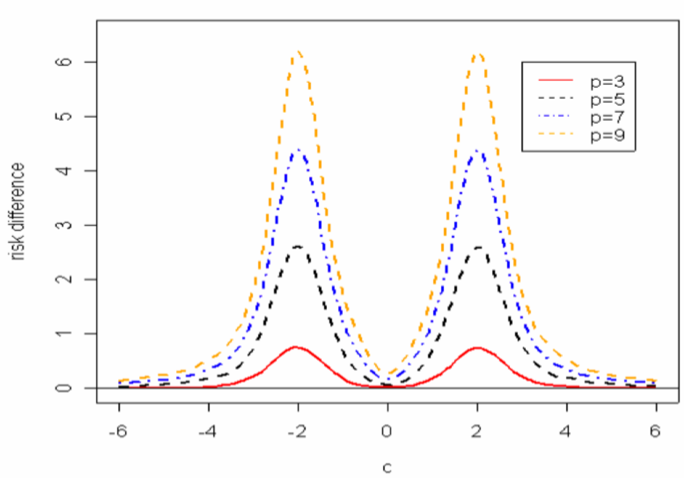
where $p_\pi^{B_i}(y \mid x)$ is a single target predictive distribution and

$$p(B_i \mid x) = \frac{w_i m_\pi^{B_i}(x; v_x)}{\sum_{i=1}^{N} w_i m_\pi^{B_i}(x; v_x)}$$

is the posterior weight on the ith prior component.

- **Theorem**: If each $m_\pi^{B_i}(z; v)$ is superharmonic, then $p_*(y \mid x)$ will dominate $p_U(y \mid x)$ and will be minimax.

- The risk reduction obtained by the multiple shrinkage predictor p_{H^*} which adaptively shrinks $p_U(y \mid x)$ towards the closer of the two points $b_1 = (2, \ldots, 2)$ and $b_2 = (-2, \ldots, -2)$ using equal weights $w_1 = w_2 = 0.5$



IV. Connecting the Estimation and Prediction Problems

IV. Connecting the Estimation and Prediction Problems

- Comparing Stein's unbiased quadratic risk expression with our unbiased KL risk expression reveals

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) = -2 \left[\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) \right]_{v=1}$$

IV. Connecting the Estimation and Prediction Problems

- Comparing Stein's unbiased quadratic risk expression with our unbiased KL risk expression reveals

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) = -2 \left[\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) \right]_{v=1}$$

- Combined with our previous KL risk difference expression reveals a fascinating connection

$$R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \left[R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) \right]_v dv$$

IV. Connecting the Estimation and Prediction Problems

- Comparing Stein's unbiased quadratic risk expression with our unbiased KL risk expression reveals

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) = -2 \left[\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) \right]_{v=1}$$

- Combined with our previous KL risk difference expression reveals a fascinating connection

$$R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \left[R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) \right]_v dv$$

- Ultimately it is this connection that yields the similar conditions for minimaxity and domination in both problems. Can we go further?

Sufficient Conditions for Admissibility

Sufficient Conditions for Admissibility

- Let $B_{KL}(\pi, q) \equiv E_\pi[R_{KL}(\mu, q)]$ be the average KL risk of $q(y \mid x)$ under π.

Sufficient Conditions for Admissibility

- Let $B_{KL}(\pi, q) \equiv E_\pi[R_{KL}(\mu, q)]$ be the average KL risk of $q(y \mid x)$ under π.
- **Theorem** (Blyth's Method): If there is a sequence of finite nonnegative measures satisfying $\pi_n(\{\mu : \|\mu\| \leq 1\}) \geq 1$ such that

$$B_{KL}(\pi_n, q) - B_{KL}(\pi_n, p_{\pi_n}) \to 0$$

then q is admissible.

Sufficient Conditions for Admissibility

- Let $B_{KL}(\pi, q) \equiv E_\pi[R_{KL}(\mu, q)]$ be the average KL risk of $q(y \mid x)$ under π.
- **Theorem** (Blyth's Method): If there is a sequence of finite nonnegative measures satisfying $\pi_n(\{\mu : \|\mu\| \leq 1\}) \geq 1$ such that

$$B_{KL}(\pi_n, q) - B_{KL}(\pi_n, p_{\pi_n}) \to 0$$

  then q is admissible.
- **Theorem**: For any two Bayes rules p_π and p_{π_n}

$$B_{KL}(\pi_n, p_\pi) - B_{KL}(\pi_n, p_{\pi_n}) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \left[B_Q(\pi_n, \hat{\mu}_\pi) - B_Q(\pi_n, \hat{\mu}_{\pi_n})\right]_v \, dv$$

  where $B_Q(\pi, \hat{\mu})$ is the average quadratic risk of $\hat{\mu}$ under π.

Sufficient Conditions for Admissibility

- Let $B_{KL}(\pi, q) \equiv E_\pi[R_{KL}(\mu, q)]$ be the average KL risk of $q(y \mid x)$ under π.
- **Theorem** (Blyth's Method): If there is a sequence of finite nonnegative measures satisfying $\pi_n(\{\mu : \|\mu\| \leq 1\}) \geq 1$ such that

$$B_{KL}(\pi_n, q) - B_{KL}(\pi_n, p_{\pi_n}) \to 0$$

  then q is admissible.
- **Theorem**: For any two Bayes rules p_π and p_{π_n}

$$B_{KL}(\pi_n, p_\pi) - B_{KL}(\pi_n, p_{\pi_n}) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \left[B_Q(\pi_n, \hat{\mu}_\pi) - B_Q(\pi_n, \hat{\mu}_{\pi_n}) \right]_v dv$$

  where $B_Q(\pi, \hat{\mu})$ is the average quadratic risk of $\hat{\mu}$ under π.
- Using the explicit construction of $\pi_n(\mu)$ from Brown and Hwang (1984), we obtain tail behavior conditions that prove admissibility of $p_U(y \mid x)$ when $p \leq 2$, and admissibility of $p_H(y \mid x)$ when $p \geq 3$.

A Complete Class Theorem

A Complete Class Theorem

- **Theorem**: In the KL risk problem, all the admissible procedures are Bayes or formal Bayes procedures.

A Complete Class Theorem

- **Theorem**: In the KL risk problem, all the admissible procedures are Bayes or formal Bayes procedures.
- Our proof uses the weak* topology from L^{∞} to L^{1} to define convergence on the action space which is the set of all proper densities on R^{p}.

A Complete Class Theorem

- **Theorem**: In the KL risk problem, all the admissible procedures are Bayes or formal Bayes procedures.
- Our proof uses the weak* topology from L^∞ to L^1 to define convergence on the action space which is the set of all proper densities on R^p.
- A Sketch of the Proof:
  1. All the admissible procedures are non-randomized.
  2. For any admissible procedure $p(\cdot \mid x)$, there exists a sequence of priors $\pi_i(\mu)$ such that $p_{\pi_i}(\cdot \mid x) \to p(\cdot \mid x)$ weak* for a.e. x.
  3. We can find a subsequence $\{\pi_{i''}\}$ and a limit prior π such that $p_{\pi_{i''}}(\cdot \mid x) \to p_\pi(\cdot \mid x)$ weak* for almost every x. Therefore, $p(\cdot \mid x) = p_\pi(\cdot \mid x)$ for a.e. x, i.e. $p(\cdot \mid x)$ is a Bayes or a formal Bayes rule.

V. Predictive Estimation for Linear Regression

V. Predictive Estimation for Linear Regression

- Observe $\quad X_{m\times 1} = A_{m\times p}\,\beta_{p\times 1} + \varepsilon_{m\times 1}$
  and predict $\quad Y_{n\times 1} = B_{n\times p}\,\beta_{p\times 1} + \tau_{n\times 1}$
  - $\varepsilon \sim N_m(0, I_m)$ is independent of $\tau \sim N_n(0, I_n)$
  - $rank(A'A) = p$

V. Predictive Estimation for Linear Regression

- Observe $X_{m\times 1} = A_{m\times p}\,\beta_{p\times 1} + \varepsilon_{m\times 1}$
  and predict $Y_{n\times 1} = B_{n\times p}\,\beta_{p\times 1} + \tau_{n\times 1}$
  - $\varepsilon \sim N_m(0, I_m)$ is independent of $\tau \sim N_n(0, I_n)$
  - $rank(A'A) = p$

- Given a prior π on β, the Bayes procedure $p_\pi^L(y \mid x)$ is

$$p_\pi^L(y \mid x) = \frac{\int p(x \mid A\beta)p(y \mid B\beta)\pi(\beta)d\beta}{\int p(x \mid A\beta)\pi(\beta)d\beta}$$

V. Predictive Estimation for Linear Regression

- Observe $\quad X_{m\times 1} = A_{m\times p}\,\beta_{p\times 1} + \varepsilon_{m\times 1}$
  and predict $\quad Y_{n\times 1} = B_{n\times p}\,\beta_{p\times 1} + \tau_{n\times 1}$
  - $\varepsilon \sim N_m(0, I_m)$ is independent of $\tau \sim N_n(0, I_n)$
  - $rank(A'A) = p$

- Given a prior π on β, the Bayes procedure $p^L_\pi(y \mid x)$ is

$$p^L_\pi(y \mid x) = \frac{\int p(x \mid A\beta) p(y \mid B\beta) \pi(\beta) d\beta}{\int p(x \mid A\beta) \pi(\beta) d\beta}$$

- The Bayes procedure $p^L_U(y \mid x)$ under the uniform prior $\pi_U \equiv 1$ is minimax with constant risk

The Key Marginal Representation

The Key Marginal Representation

- For any prior π,

$$p^L_\pi(y \mid x) = \frac{m_\pi(\hat{\beta}_{x,y}, (C'C)^{-1})}{m_\pi(\hat{\beta}_x, (A'A)^{-1})}\, p^L_U(y \mid x)$$

where $C_{(m+n)\times p} = (A', B')'$ and

$$\hat{\beta}_x = (A'A)^{-1}A'x \sim N_p(\beta, (A'A)^{-1})$$

$$\hat{\beta}_{x,y} = (C'C)^{-1}C'(x', y')' \sim N_p(\beta, (C'C)^{-1})$$

Risk Improvement over $p_U^L(y \mid x)$

Risk Improvement over $p_U^L(y \mid x)$

- Here the difference between the KL risks of $p_U^L(y \mid x)$ and $p_\pi^L(y \mid x)$ can be expressed as

$$R_{KL}(\beta, p_U^L) - R_{KL}(\beta, p_\pi^L) =$$
$$E_{\beta,(C'C)^{-1}} \log m_\pi(\hat{\beta}_{x,y}; (C'C)^{-1}) - E_{\beta,(A'A)^{-1}} \log m_\pi(\hat{\beta}_x; (A'A)^{-1})$$

Risk Improvement over $p^L_U(y \mid x)$

- Here the difference between the KL risks of $p^L_U(y \mid x)$ and $p^L_\pi(y \mid x)$ can be expressed as

$$R_{KL}(\beta, p^L_U) - R_{KL}(\beta, p^L_\pi) =$$
$$E_{\beta,(C'C)^{-1}} \log m_\pi(\hat{\beta}_{x,y}; (C'C)^{-1}) - E_{\beta,(A'A)^{-1}} \log m_\pi(\hat{\beta}_x; (A'A)^{-1})$$

- Minimaxity of $p^L_\pi(y \mid x)$ is here obtained when

$$\frac{\partial}{\partial \omega} E_{\mu, V_\omega} \log m_\pi(Z; V_\omega) < 0$$

where

$$V_\omega \equiv \omega(A'A)^{-1} + (1-\omega)(C'C)^{-1}$$

Risk Improvement over $p_U^L(y \mid x)$

- Here the difference between the KL risks of $p_U^L(y \mid x)$ and $p_\pi^L(y \mid x)$ can be expressed as

$$R_{KL}(\beta, p_U^L) - R_{KL}(\beta, p_\pi^L) =$$
$$E_{\beta,(C'C)^{-1}} \log m_\pi(\hat{\beta}_{x,y}; (C'C)^{-1}) - E_{\beta,(A'A)^{-1}} \log m_\pi(\hat{\beta}_x; (A'A)^{-1})$$

- Minimaxity of $p_\pi^L(y \mid x)$ is here obtained when

$$\frac{\partial}{\partial \omega} E_{\mu, V_\omega} \log m_\pi(Z; V_\omega) < 0$$

where

$$V_\omega \equiv \omega(A'A)^{-1} + (1-\omega)(C'C)^{-1}$$

- This leads to weighted superharmonic conditions on m_π and π for minimaxity.

Some References

George, E.I., Liang, F. and Xu, X. (2006). Improved Minimax Predictive Densities Under Kullback-Leibler Loss. *Annals of Statistics*, 34 1 78–91.

Brown, L.D., George, E.I., and Xu, X. (2008). Admissible Predictive Density Estimation. *Annals of Statistics*, 36, 3, 1156–1170.

George, E.I. and Xu, X. (2008). Predictive Density Estimation for Multiple Regression. *Econometric Theory*, 24, 528–544.

George, E.I., Liang, F. and Xu, X. (2012). From Minimax Shrinkage Estimation to Minimax Shrinkage Prediction. *Statistical Science*, Vol. 27, No. 1 82–94.

Mukherjee, G. and Johnstone, I.M. (2018). Exact Minimax Estimation of the Predictive Density in Sparse Gaussian Models. *Annals of Statistics*.

Thank you!