Econ 508B: Lecture 7

Optimal Data Reduction

Hongyi Liu

Washington University in St. Louis

August 4, 2017

Outline

Outline

Retrospect

Last time, we have set out to develop optimal inference procedures and, data reduction and relevant assertions:

- not all data is relevant;
- irrelevant data may increase risk;
- a notion of lossless data compression (**sufficiency**) showed that data reduction via sufficiency never hurts;

Question: How much data can we throw away via data reduction?

Retrospect

Last time, we have set out to develop optimal inference procedures and, data reduction and relevant assertions:

- not all data is relevant;
- irrelevant data may increase risk;
- a notion of lossless data compression (**sufficiency**) showed that data reduction via sufficiency never hurts;

Question: How much data can we throw away via data reduction?

- The answer to this question is about **optimal data reduction**.
- To solve this question, we will establish some concepts of **Minimal Sufficiency** and **Completeness**. Before doing so, we firstly introduce an especially important class of models known as exponential family models with an amount of nice properties.

Exponential Families

Definition 1.1

A model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ forms an $s-$ dimensional exponential family if each P_θ has a density of the form

$$p(x;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x).$$

Each component has a corresponding name:

- $\eta_i(\theta) \in \mathbb{R}$: natural parameters.
- $T_i(x) \in \mathbb{R}$: sufficient statistics.
- $B(\theta) \in \mathbb{R}$: log-partition function, responsible for normalization of $p(x;\theta)$, which depends on θ.
- $h(x) \in \mathbb{R}$: base measure.

Example: Normal Distribution with unknown mean

Consider $X \sim N(\mu, \sigma^2)$ and the density of X is

$$\begin{aligned} f(x|\sigma) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} I_{x\in\mathbb{R}} \\ &= \frac{1}{\sqrt{2\pi}} e^{(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma)} I_{x\in\mathbb{R}} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \exp\left(\eta(\theta)^\top T(x) - (\underbrace{\log\sigma + \frac{\mu^2}{2\sigma^2}}_{B(\theta)}) \right) \end{aligned}$$

where $T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}, B(\theta) = \log\sigma + \frac{\mu^2}{2\sigma^2}, \theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$.

Canonical Form: simpler

An exponential family is in **canonical form** when the density has the form

$$p(x;\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x)$$

This parameterizes the density in terms of the **natural parameters** η instead of θ.

- For a given base measure h and collection of sufficient statistics $\{T_i(x)\}$, only some values of η will give rise to valid, normalizable densities.
- The set of all valid natural parameters is called the **natural parameter space:**

$$\Theta = \{\eta : 0 < \int \exp(\sum_{i=1}^{s} \eta_i T_i(x)) h(x) d\mu(x) < \infty\}.$$

- Θ is a convex subspace of $\mathbb{R}^n$.

Reducing the dimension

There are two cases when the superficial dimension of an s-dimensional exponential family can be reduced.

- CASE 1: $T_i(x)$s satisfy with a *linear constraint/**unidentifiability***.

Example 1.1 ($X \sim \exp(\eta_1, \eta_2)$)

$p(x; \eta_1, \eta_2) = \exp(-\eta_1 x - \eta_2 x + \log(\eta_1 + \eta_2)) I_{[0,\infty)}(x), T_1 = T_2 = x$, i.e., they are linear dependent and we could collapse (η_1, η_2) to $\eta = \eta_1 + \eta_2$.

Definition 1.2 (unidentifiability)

If $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, then θ is **unidentifiable** if for two parameters $\theta_1 \neq \theta_2, P_{\theta_1} = P_{\theta_2}$.

- CASE2: η_i's satisfy a linear constraint.

Example 1.2

$p(x; \eta) \propto \exp(\eta_1 x + \eta_2 x^2) = \exp(\eta_1 (x - x^2) + x^2)$ for $\eta_1 + \eta_2 = 1$

Continued

When neither of the above cases holds, we call an exponential family **minimal**.

Definition 1.3

A canonical exponential family $\mathcal{P} = \{P_\eta : \eta \in H\}$ is **minimal** if

- $\sum_{i=1}^{s} \lambda_i T_i(x) = \lambda_0, \forall x \in \mathcal{X} \Rightarrow \lambda_i = 0, \forall i \in \{1, ..., s\}$ (no linear T_i constraints).
- $\sum_{i=1}^{s} \lambda_i \eta_i = \lambda_0, \forall \eta \in H \Rightarrow \lambda_i = 0, \forall i \in \{1, ..., s\}$ (no linear η_i constraints).

Definition 1.4

Suppose $\mathcal{P} = \{P_\eta : \eta \in H\}$ is an s-dimensional minimal exponential family. If H contains an open s-dimensional rectangle, then $\mathcal{P}$ is called **full rank**. Otherwise, $\mathcal{P}$ is **curved**. In curved exponential families, the η_i's are related in a non-linear way.

- **Non-minimal:**(the dimension can be reduced), e.g., $\mathbb{N}(\sigma, \sigma)$.
- **Minimal & Curved:**e.g., $\mathbb{N}(\sigma, \sigma^2)$.
- **Minimal & Full-Rank:** e.g., $\mathbb{N}(\mu, \sigma^2)$.

Properties of Exponential Families

- Exponential family data is **exceptionally compressible** by Neyman-Fisher factorization criterion: there exists an s-dimensional sufficient statistic for any sample size.

Example 1.3

If i.i.d. $X_1, ..., X_n \sim p(x;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$, then

$$p(x_1, ..., x_n;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \sum_{j=1}^{n} T_i(x_j) - nB(\theta)\right) \prod_{j=1}^{n} h(x_j).$$

By NFFC, $\left(\sum_{j=1} T_1(x_j), ..., \sum_{j=1} T_s(x_j)\right)$ is therefore a **sufficient statistic.**

- Any expectation of continuous and integrable function is **infinitely differentiable**.

Outline

Minimal

Now, we return to our initial question of optimal data reduction. We begin by defining a function of the data that cannot be reduced without sacrificing information about the model.

Definition 2.1

A sufficient statistic T is **minimal** if for every sufficient statistic T', T is a function of T'. Equivalently, T is minimal if for every sufficient statistic $T', T(x) = T(y)$ whenever $T'(x) = T'(y)$.

Minimal Sufficiency

Definition 2.2

A sufficient statistic T is **minimal sufficient** if for any other sufficient statistic T' and every $x, y \in \mathcal{X}, T(x) = T(y)$ whenever $T'(x) = T'(y)$.
In other words, T is a function of T', i.e., $\exists f$ such that $T(x) = f(T'(x))$ for any $x \in \mathcal{X}$.

A sufficient condition

The definition of minimal sufficiency is a bit abstract to deal with. Here provides a sufficient condition for minimal sufficiency.

Theorem 2.1

Let $\{p(x;\theta), \theta \in \Omega\}$ be a family of densities (w.r.t some measure μ). Suppose that there exists T such that for any $x, y \in \mathcal{X}$,

$$\frac{p(x;\theta)}{p(y;\theta)} \text{ does not depend on } \theta \Leftrightarrow T(x) = T(y).$$

Then T is a minimal sufficient statistic.

Example

Example 2.1

Let $X_1, ..., X_n \overset{i.i.d}{\sim} N(\sigma, \sigma^2), \theta = \sigma > 0, x, y \in \mathcal{X}$.

$$\frac{p(x;\theta)}{p(y;\theta)} = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_i x_i^2 + \frac{\sigma}{\sigma^2}\sum_i x_i - \frac{n\sigma^2}{2\sigma^2}\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_i y_i^2 + \frac{\sigma}{\sigma^2}\sum_i y_i - \frac{n\sigma^2}{2\sigma^2}\right)}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{1}{\sigma}\left(\sum_i x_i - \sum_i y_i\right)\right)$$

Then $T(X) = (T_1(X), T_2(X)) = (\sum_i X_i^2, \sum_i X_i)$ is minimal sufficient.

Proof: "$\Leftarrow$", if $T(x) = T(y)$ then the ratio is equal to 1 and hence does not depend on θ;
"$\Rightarrow$", if for some $x, y \in \mathbb{R}$, the ratio does not depend on θ which leads to

$$-\frac{1}{2\sigma^2}\left(T_1(x) - T_1(y)\right) + \frac{1}{\sigma}\left(T_2(x) - T_2(y)\right) = 0, \forall \sigma > 0,$$

Therefore,

$$T_1(x) - T_1(y) = 2\sigma(T_2(x) - T_2(y)), \forall \sigma > 0.$$

It suffices to hold this equation for all $\sigma > 0$ only when $T_1(x) = T_1(y)$ and $T_2(x) = T_2(y)$.
Consequently, T is a **minimal sufficient statistic**.

Outline

Complete and Ancillary

Definition 3.1 (Completeness)

Suppose that $T = T(X)$ is a statistic for $X \sim \mathbb{P}_\theta \in \mathcal{P}$. Then T is a **complete** statistic for θ if for any function $r : T \to \mathbb{R}$

$$\mathbb{E}_\theta[r(T)] = 0 \text{ for all } \theta \in \Theta \Rightarrow \mathbb{P}_\theta[r(T) = 0] = 1 \text{ for all } \theta \in \Theta.$$

It is closely related to the idea of **identifiability**.

Definition 3.2 (Ancillary)

Consider $V = V(X)$ is a statistic for $X \sim \mathbb{P}_\theta \in \mathcal{P}$. If the distribution of V does not depend on θ, then V is **ancillary** statistic for θ.

Remark 3.1

The notion of an ancillary statistic is complementary to the notion of a sufficient statistic, which contains all available information about θ while an ancillary statistic contains no information about θ.

Basu's Theorem

Theorem 3.1 (Basu's Theorem)

Suppose that T is is complete and sufficient for a parameter θ and that V is an ancillary statistic for θ. Then T and V are **independent**.

Example 3.1 (normal distribution)

Let $X_1, ..., X_n$ be i.i.d. normal variables with mean μ and variance σ^2. Then with respect to the parameter μ, one can have the following **sample mean** and **sample variance**

$$\hat{\mu} = \frac{\sum X_i}{n}, \hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

and show that $\hat{\mu}$ is a complete sufficient statistic, i.e. it is all the information one can derive to estimate μ, and no more $\hat{\sigma}^2$ is an ancillary statistic – its distribution does not depend on μ.

Outline

From Data Reduction to Risk Reduction

Question: How does optimal data compression translate into better decision procedures?

To answer this question, we turn our attention on point estimation and show that under some condition, the data reduction can be used to improve an unbiased estimator.

Rao-Blackwell Theorem

Theorem 4.1

Let T be sufficient for $\mathbb{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, that $\delta(X)$ is an estimator for $g(\theta)$ for which $\mathbb{E}(\delta(X))$ exists, and that $R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X)) < \infty$. If $L(\theta, \cdot)$ is convex (as a function of $d \in \mathcal{D}$), then by *Jensen's inequality*

$$R(\theta, \eta) \leq R(\theta, \delta) \text{ for } \eta(T(X)) = \mathbb{E}(\delta(X)|T(X))$$

If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \eta) < R(\theta, \delta)$ for any θ unless $\eta = \delta$ with probability 1.

Example

Let $X_1, ..., X_n \overset{iid}{\sim} Ber(\theta)$ for $\theta \in (0, 1)$, and consider the loss function $L(\theta, d) = (\theta - d)^2$.

- Suppose we begin with an estimator $\delta(X) = X_1$ that only makes use of the first data point.
- We know that $T(X) = \bar{X}$ is a sufficient statistic as proved in the previous slide and apply the *Rao-Blackwell theorem* to improve our estimator δ as follows:

$$\eta(T(X)) = \mathbb{E}_\theta(\delta(X)|T(X)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_1|\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i|\bar{X}) = \bar{X}.$$

- Regarding risk function, $R(\theta, \eta) = \frac{\theta(1-\theta)}{n} < \theta(1-\theta) = R(\theta, \delta)$.

Outline

Motivation: unbiased estimation

Recall that an estimator is **unbiased** if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$.

- Although it is hard to find a uniformly minimal risk for unconstrained estimator, we can find an unbiased estimator with uniformly minimal risk, namely an unbiased δ satisfying $R(\theta, \delta) \leq R(\theta, \delta'), \forall \theta$ and for all unbiased estimators δ'.
- An estimator satisfying the above conditions is referred to a **uniformly minimum risk unbiased estimator(UMRUE)**

Uniformly Minimum Variance Unbiased Estimator

When $L(\theta, d) = (\theta - d)^2$, $UMRUE$ is also called an **uniformly minimum variance unbiased estimator(UMVUE)**.

- using the decomposition of mean squared error

$$\mathbb{E}_\theta = [\theta, \delta(X)^2] = \underbrace{(\mathbb{E}_\theta[(\delta(X)] - \theta))^2}_{Bias^2} + \underbrace{\mathbb{E}[(\delta(X) - \mathbb{E}_\theta[\delta(X)])^2]}_{Variance}$$

- If the bias is restricted to be 0, risk minimization problem is reduced into minimizing variance.
- In particular, in *Econometrics*, $Bias = 0$ is similar to **exogenous condition** and minimizing variance is the analog of OLS.

Lehmann-Scheffe Theorem

- The next result shows the importance of statistics that are both complete and sufficient and provide a connection between risk reduction of unbiased estimator and data reduction when the loss function is convex.

Theorem 5.1

Suppose that T is sufficient and complete for θ and that $V = r(T(X))$ is an unbiased estimator of a real parameter $g = g(\theta)$. Then $V = r(T(X))$ is

- the only unbiased function of $T(X)$,
- an UMRUE under any convex loss function,
- the unique UMRUE under any strictly convex loss function,
- the unique UMVUE.

Optimal Unbiased Estimation

Lehmann-Scheffe Theorem gives rise to several useful strategies for finding UMRUE's under convex function.

Strategies for optimal unbiased estimation under convex function

(1) **Rao Blackwell Theorem**
   (i) find any complete sufficient statistics $T(X)$,
   (ii) find any unbiased estimator δ,
   (iii) Compute $\mathbb{E}[\delta(X)|T(X)]$.

(2) **Calculate** the (unique) unbiased δ fulfilled with $\mathbb{E}[\delta(X)] = g(\theta)$.

(3) **Conditioning** any unbiased function on $T(X)$.