Lecture Notes 13
The Bootstrap

See Chapter 8.

1 Introduction

Can we estimate the mean of a distribution without using a parametric model? Yes. The key idea is to first estimate the distribution function nonparametrically. Then we can get an estimate of the mean (and many other parameters) from the distribution function.

How can we get the standard error of that estimator? How do we get a confidence interval? The answer is: the bootstrap. The bootstrap is a nonparametric method for finding standard errors and confidence intervals.

Notation. Let F be a distribution function. Let p denote the probability mass function if F is discrete and the probability density function if F is continuous. The integral $\int g(x)dF(x)$ is interpreted as follows:

$$\int g(x)dF(x) = \begin{cases} \sum_j g(x_j)p(x_j) & \text{if } F \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } F \text{ is continuous.} \end{cases} \tag{1}$$

For $0 < \alpha < 1$ define z_α by $\mathbb{P}(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$. Thus $z_\alpha = \Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$.

2 Review of The Empirical Distribution Function

The bootstrap uses the empirical distribution function. Let $X_1, \ldots, X_n \sim F$ where $F(x) = \mathbb{P}(X \leq x)$ is a distribution function on the real line. We can estimate F with the **empirical distribution function** $\widehat{F}_n$, the cdf that puts mass $1/n$ at each data point X_i.

Recall that the empirical distribution function $\widehat{F}_n$ is defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{2}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases} \tag{3}$$

According to the **Glivenko–Cantelli Theorem**,

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{as}} 0. \tag{4}$$

Hence, $\widehat{F}_n$ is a consistent estimator of F. In fact, the convergence is fast. According to the **Dvoretzky–Kiefer–Wolfowitz (DKW)** inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{5}$$

If $\epsilon_n = c_n/\sqrt{n}$ where $c_n \to \infty$, then $\mathbb{P}(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon_n) \to 0$. Hence, $\sup_x |F(x) - \widehat{F}_n(x)| = O_P(n^{-1/2})$. One last point: from (1) it follows that for any function g, $\int g(x)d\widehat{F}_n(x) = n^{-1}\sum_{i=1}^n g(X_i)$.

3 Statistical Functionals

Recall that a **statistical functional** $T(F)$ is any function of of the cdf F. Examples include the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$, $m = F^{-1}(1/2)$, and the largest eigenvalue of the covariance matrix Σ.

The **plug-in estimator** of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n). \tag{6}$$

Let $\widehat{\mathsf{se}}$ be an estimate of the standard error of $T(\widehat{F}_n)$. (We will see how to get this later.) In many cases, it turns out that

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2). \tag{7}$$

In that case, an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$T(\widehat{F}_n) \pm z_{\alpha/2}\, \widehat{\mathsf{se}}. \tag{8}$$

Example 1 (The mean) *Let $\mu = T(F) = \int x\, dF(x)$. The plug-in estimator is $\widehat{\mu} = \int x\, d\widehat{F}_n(x) = \overline{X}_n$. The standard error is $\mathsf{se} = \sqrt{\mathsf{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. If $\widehat{\sigma}$ denotes an estimate of σ, then the estimated standard error is $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. A Normal-based confidence interval for μ is $\overline{X}_n \pm z_{\alpha/2}\, \widehat{\sigma}/\sqrt{n}$.*

Example 2 *A functional of the form $\int a(x)dF(x)$ is called a* **linear functional.** *(Recall that $\int a(x)dF(x)$ is defined to be $\int a(x)p(x)dx$ in the continuous case and $\sum_j a(x_j)p(x_j)$ in the discrete case.) The empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each X_i. Hence, if $T(F) = \int a(x)dF(x)$ is a linear functional then the plug-in estimator for linear functional $T(F) = \int a(x)dF(x)$ is:*

$$T(\widehat{F}_n) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n a(X_i). \tag{9}$$

Example 3 (The variance) *Let* $\sigma^2 = \mathsf{Var}(X) = \int x^2\, dF(x) - \left(\int x\, dF(x)\right)^2$. *The plug-in estimator is*

$$\widehat{\sigma}^2 = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 \tag{10}$$

$$= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \tag{11}$$

$$= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \tag{12}$$

Example 4 (The skewness) *Let* μ *and* σ^2 *denote the mean and variance of a random variable* X. *The skewness — which measures the lack of symmetry of a distribution — is defined to be*

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}. \tag{13}$$

To find the plug-in estimate, first recall that $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$ *and* $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$. *The plug-in estimate of* κ *is*

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}. \tag{14}$$

Example 5 (Correlation) *Let* $Z = (X, Y)$ *and let* $\rho = T(F) = \mathbb{E}(X-\mu_X)(Y-\mu_Y)/(\sigma_x\sigma_y)$ *denote the correlation between* X *and* Y, *where* $F(x,y)$ *is bivariate. We can write* $T(F) = a(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$ *where*

$$\begin{array}{lll} T_1(F) = \int x\, dF(z) & T_2(F) = \int y\, dF(z) & T_3(F) = \int xy\, dF(z) \\ T_4(F) = \int x^2\, dF(z) & T_5(F) = \int y^2\, dF(z) & \end{array} \tag{15}$$

and

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}. \tag{16}$$

Replace F *with* $\widehat{F}_n$ *in* $T_1(F)$, $\ldots$, $T_5(F)$, *and take*

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)). \tag{17}$$

We get

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}} \tag{18}$$

which is called the **sample correlation**.

Example 6 (Quantiles) *Let* F *be strictly increasing with density* f. *Let* $T(F) = F^{-1}(p)$ *be the* p^{th} *quantile. The estimate of* $T(F)$ *is* $\widehat{F}_n^{-1}(p)$. *We have to be a bit careful since* $\widehat{F}_n$ *is not invertible. To avoid ambiguity we define* $\widehat{F}_n^{-1}(p) = \inf\{x :\ \widehat{F}_n(x) \geq p\}$. *We call* $\widehat{F}_n^{-1}(p)$ *the* p^{th} **sample quantile**.

4 The Bootstrap

Let $T_n = g(X_1, \ldots, X_n)$ be a statistic and let $\mathsf{Var}_F(T_n)$ denote the variance of T_n. We have added the subscript F to emphasize that the variance is itself a function of F. In other words

$$\mathsf{Var}_F(T_n) = \int\int \cdots \int (g(X_1, \ldots, X_n) - \mu)^2 dF(x_1) dF(x_2) \cdots dF(x_n)$$

where

$$\mu = \mathbb{E}(T_n) = \int\int \cdots \int g(X_1, \ldots, X_n) dF(x_1) dF(x_2) \cdots dF(x_n).$$

If we knew F we could, at least in principle, compute the variance. For example, if $T_n = n^{-1} \sum_{i=1}^n X_i$, then

$$\mathsf{Var}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}. \tag{19}$$

In other words, the variance of $\widehat{\theta} = T(F_n)$ is itself a function of F. We can write

$$\mathsf{Var}_F(T_n) = U(F)$$

for some U. Therefore, to estimate $\mathsf{Var}_F(T_n)$ we can use

$$\widehat{\mathsf{Var}_F(T_n)} = U(\widehat{F}_n).$$

This is the bootstrap estimate of the standard error. To repeat: we estimate $U(F) = \mathsf{Var}_F(T_n)$ with $U(\widehat{F}_n) = \mathsf{Var}_{\widehat{F}_n}(T_n)$. In other words, we use a plug-in estimator of the variance.

But how can we compute $\mathsf{Var}_{\widehat{F}_n}(T_n)$? We approximate it with a simulation estimate denoted by v_{boot}. Specifically, we do the following steps:

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$\widehat{v} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2. \tag{20}$$

By the law of large numbers, $\widehat{v} \xrightarrow{\text{as}} \mathsf{Var}_{\widehat{F}_n}(T_n)$ as $B \to \infty$. The estimated standard error of T_n is $\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{\widehat{v}}$. The following diagram illustrates the bootstrap idea:

$$\begin{array}{rccccc} \text{Real world:} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\ \text{Bootstrap world:} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

Bootstrap for the Median

```
Given data X = (X(1), ..., X(n)):

T     = median(X)
Tboot = vector of length B
for(i in 1:N){
    Xstar    = sample of size n from X (with replacement)
    Tboot[i] = median(Xstar)
    }
se = sqrt(variance(Tboot))
```

Figure 1: Pseudo-code for bootstrapping the median.

$$\mathsf{Var}_F(T_n) \overbrace{\approx}^{O(1/\sqrt{n})} \mathsf{Var}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{O(1/\sqrt{B})} \widehat{v}. \tag{21}$$

How do we simulate from $\widehat{F}_n$? Since $\widehat{F}_n$ gives probability $1/n$ to each data point, drawing n points at random from $\widehat{F}_n$ is the same as drawing a sample of size n with replacement from the original data. Therefore step 1 can be replaced by:

1. **Draw $X_1^*, \ldots, X_n^*$ with replacement from $X_1, \ldots, X_n$.**

Example 7 *Figure 1 shows pseudo-code for using the bootstrap to estimate the standard error of the median.*

5 The Parametric Bootstrap

So far, we have estimated F nonparametrically. There is also a **parametric bootstrap.** If F_θ depends on a parameter θ and $\widehat{\theta}$ is an estimate of θ, then we simply sample from $F_{\widehat{\theta}}$ instead of $\widehat{F}_n$. This is just as accurate, but much simpler than, the delta method. Here is more detail.

Suppose that $X_1, \ldots, X_n \sim p(x; \theta)$. Let $\widehat{\theta}$ be the mle. Let $\tau = g(\theta)$. Then $\widehat{\tau} = g(\widehat{\theta})$. To get the standard error of $\widehat{\tau}$ we need to compute the Fisher information and then do the delta method. The bootstrap allows us to avoid both steps. We just do the following:

1. Compute the estimate $\widehat{\theta}$ from the data $X_1, \ldots, X_n$.
2. Draw a sample $X_1^*, \ldots, X_n^* \sim p(x; \widehat{\theta})$. Compute $\widehat{\theta}_1^*$ and $\widehat{\tau}_1^* = g(\widehat{\theta}_1^*)$ from the new data. Repeat B times to get $\widehat{\tau}_1^*, \ldots, \widehat{\tau}_B^*$.

3. Compute the standard deviation

$$\widehat{\text{se}} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\tau}_j^* - \overline{\tau})^2 \qquad \text{where} \qquad \overline{\tau} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_j^*. \tag{22}$$

No need to get the Fisher information or do the delta method.

6 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. They vary in ease of calculation and accuracy. We discuss two of them.

Normal Interval. The simplest is the Normal interval

$$\widehat{\theta}_n \pm z_{\alpha/2}\ \widehat{\mathsf{se}}_{\text{boot}} \tag{23}$$

where $\widehat{\mathsf{se}}_{\text{boot}}$ is the bootstrap estimate of the standard error.

Pivotal Intervals. Generally, this is the preferred method. Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$. We can construct an approximate confidence interval for θ using the (approximate) pivot $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ as follows.

First, define

$$G_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta} - \theta) \leq t).$$

The distribution of G_n is not known. Suppose, for the moment, that we did know G_n. Let

$$z_{\alpha/2} = G_n^{-1}(\alpha/2), \quad z_{1-\alpha/2} = G_n^{-1}(1-\alpha/2).$$

Define

$$C_n = \left[\widehat{\theta} - \frac{z_{1-\alpha/2}}{\sqrt{n}},\ \ \widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n}}\right].$$

Then

$$\begin{aligned}
\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(\widehat{\theta} - \frac{z_{1-\alpha/2}}{\sqrt{n}} \leq \theta \leq \widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n}}\right) \\
&= \mathbb{P}(z_{\alpha/2} \leq \sqrt{n}(\widehat{\theta} - \theta) \leq z_{1-\alpha/2}) \\
&= G_n(z_{1-\alpha/2}) - G_n(z_{\alpha/2}) \\
&= (1-\alpha/2) - (\alpha/2) = 1-\alpha.
\end{aligned}$$

The idea is to estimate the distribution G_n. The estimate is

$$\begin{aligned}
\widehat{G}_n(t) &= \mathbb{P}(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) \leq t \,|\, X_1, \ldots, X_n) \\
&\approx \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}) \leq t).
\end{aligned}$$

Let

$$\widehat{z}_{\alpha/2} = G_n^{-1}(\alpha/2), \quad \widehat{z}_{1-\alpha/2} = G_n^{-1}(1-\alpha/2).$$

Define

$$C_n = \left[\widehat{\theta} - \frac{\widehat{z}_{1-\alpha/2}}{\sqrt{n}}, \;\; \widehat{\theta} - \frac{\widehat{z}_{\alpha/2}}{\sqrt{n}}\right].$$

If

$$\sup_t |\widehat{G}_n(t) - G_n(t)| \xrightarrow{P} 0$$

then $\widehat{z}_{\alpha/2} - z_{\alpha/2} \xrightarrow{P} 0$ and $\widehat{z}_{1-\alpha/2} - z_{1-\alpha/2} \xrightarrow{P} 0$. Then

$$\begin{aligned}
\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(\widehat{\theta} - \frac{\widehat{z}_{1-\alpha/2}}{\sqrt{n}} \leq \theta \leq \widehat{\theta} - \frac{\widehat{z}_{\alpha/2}}{\sqrt{n}}\right) \\
&= \mathbb{P}(\widehat{z}_{\alpha/2} \leq \sqrt{n}(\widehat{\theta} - \theta) \leq \widehat{z}_{1-\alpha/2}) \\
&= G_n(\widehat{z}_{1-\alpha/2}) - G_n(\widehat{z}_{\alpha/2}) \\
&= G_n(z_{1-\alpha/2}) - G_n(z_{\alpha/2}) + o_P(1) \\
&= (1-\alpha/2) - (\alpha/2) + o_P(1) = 1 - \alpha + o_P(1).
\end{aligned}$$

The key question is this: how do we prove that

$$\sup_t |\widehat{G}_n(t) - G_n(t)| \xrightarrow{P} 0?$$

The idea is explained in the following diagram:

$$\begin{array}{rcl}
\sqrt{n}(\widehat{\theta} - \theta) & \approx & N(0, \tau^2) \\
 & & \wr\wr \\
\sqrt{n}(\widehat{\theta}^* - \theta)|X_1, \ldots, X_n & \approx & N(0, \widehat{\tau}^2)
\end{array}$$

Let's make this formal in the simple case of the sample mean. First, let us recall a Theorem from Lecture 4.

Theorem 8 (Berry-Esseen Theorem) *Let $X_1, \ldots, X_n \sim P$. Let $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i]$. Assume that $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. Let*

$$F_n(z) = \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z\right).$$

Then

$$\sup_z |F_n(z) - \Phi(z)| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}.$$

Theorem 9 *Let $X_1, \ldots, X_n \sim F$ and let $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i]$. Assume that $\mu_3 < \infty$. Let*

$$G_n(t) = \mathbb{P}(\sqrt{n}(\overline{X} - \mu) \leq t)$$
$$\widehat{G}_n(t) = \mathbb{P}(\sqrt{n}(\overline{X}^* - \overline{X}) \leq t \,|\, X_1, \ldots, X_n).$$

Then

$$\sup_t |\widehat{G}_n(t) - G_n(t)| \xrightarrow{P} 0.$$

Proof. Let Φ_σ denote a Normal cdf with mean 0 and variance σ^2. From the Berry-Esseen theorem

$$\begin{aligned}\sup_t |G_n(t) - \Phi_\sigma(t)| &= \sup_t |\mathbb{P}(\sqrt{n}(\overline{X} - \mu) \leq t) - \Phi_\sigma(t)| \\ &\leq \frac{C\mu_3}{\sigma^3\sqrt{n}} = O(1/\sqrt{n}).\end{aligned}$$

Now $\overline{X}^*$ is based on an iid sample from F_n which has mean $\overline{X}_n$ and variance s_n^2. So, conditional on $(X_1, \ldots, X_n)$,

$$\begin{aligned}\sup_t |\widehat{G}_n(t) - \Phi_s(t)| &= \sup_t |\mathbb{P}(\sqrt{n}(\overline{X}^* - \overline{X}) \leq t | X_1, \ldots, X_n) - \Phi_s(t)| \\ &\leq \frac{C\widehat{\mu}_3}{s^3\sqrt{n}}.\end{aligned}$$

Since $\widehat{\mu}_3 \xrightarrow{P} \mu_3$ and $s^3 \xrightarrow{P} \sigma_3$, we have that $\sup_t |\widehat{G}_n(t) - \Phi_s(t)| = O_P(1/\sqrt{n})$. We also know that $s - \sigma = O_P(1/\sqrt{n})$. This implies that

$$\sup_t |\Phi_s(t) - \Phi_\sigma(t)| = O_P(1/\sqrt{n}).$$

Finally, using the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{G}_n(t) - G_n(t)| &\leq \sup_t |\widehat{G}_n(t) - \Phi_s(t)| + \sup_t |\Phi_s(t) - \Phi_\sigma(t)| + \sup_t |G_n(t) - \Phi_\sigma(t)| \\ &= O_P(1/\sqrt{n}).\end{aligned}$$

■

The extension to more general parameters is quite involved but the idea is basically the same. Note that we have actually proved that the coverage of the bootstrap interval is $1 - \alpha + O(1/\sqrt{n})$.

7 Remarks About The Bootstrap

1. The bootstrap is nonparametric but it does require some assumptions. You can't assume it is always valid.

2. The bootstrap is an asymptotic method. Thus the coverage of the confidence interval is $1-\alpha+r_n$ where the remainder $r_n \to 0$ as $n \to \infty$.
3. There is a related method called the jackknife where the standard error is estimated by leaving out one observation at a time. However, the bootstrap is valid under weaker conditions than the jackknife. See Shao and Tu (1995).
4. Another way to construct a bootstrap confidence interval is to set $C = [a, b]$ where a is the $\alpha/2$ quantile of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ and b is the $1-\alpha/2$ quantile. This is called the percentile interval. This interval seems very intuitive but does not have theoretical support.
5. There are many cases where the bootstrap is not formally justified. This is especially true with discrete structures like trees and graphs. Nonethless, the bootstrap can be used in an informal way to get some intuition of the variability of the procedure. But keep in mind that the formal guarantees may not apply in these cases. For example, see Holmes (2003) for a discussion of the bootstrap applied to phylogenetic tres.
6. There is a version of the bootstrap called subsampling. In this case, we draw samples of size $m < n$ without replacement. Subsampling produces valid confidence intervals under weaker conditions than the bootstrap. See Politis, Romano and Wolf (1999).
7. There are many modifications of the bootstrap that lead to more accurate confidence intervals; see Efron (1996).

8 Examples

Example 10 (The Mean of a χ^2) *The top left plot of Figure 2 shows the density for a χ^2_5 distribution. The top right plot shows a histogam of $n = 100$ draws from this distribution. Let $\theta = T(P)$ be the mean. The true value is $\theta = 5$. I simulated manu data sets. Each time, I compute a 95 percent bootstrap confidence interval. See the bottom left plot. The empirical coverage was 0.92.*

Example 11 (Estimating Eigenvalues) *Let $X_1, \ldots, X_n$ be random vectors where $X_i \in \mathbb{R}^p$ and let Σ be the covariance matrix of X_i. A common dimension reduction technique is principal components which involves finding the spectral decomposition $\Sigma = E\Lambda E^T$ where the columns of E are the eigenvectors of Σ and Λ is a diagonal matrix whose diagonal elements are the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. The data dimension can be reduced to $q < p$ by projecting each data point onto the first q eigenvalues. We choose q such that $\sum_{j=q+1}^p \lambda_j^2$ is small. Of course, we need to estimate the eigenvectors and eigenvalues. For now, let us focus on estimating the largest eigenvalue and denote this by θ. An estimate of θ is the largest principal component $\widehat{\theta}$ of the sample covariance matrix*

$$S = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^T. \tag{24}$$

It is not at all obvious how can can estimate the standard error of $\widehat{\theta}$ or how to find a confidence interval for θ. In this example, the bootstrap works as follows. Draw a sample of size n with

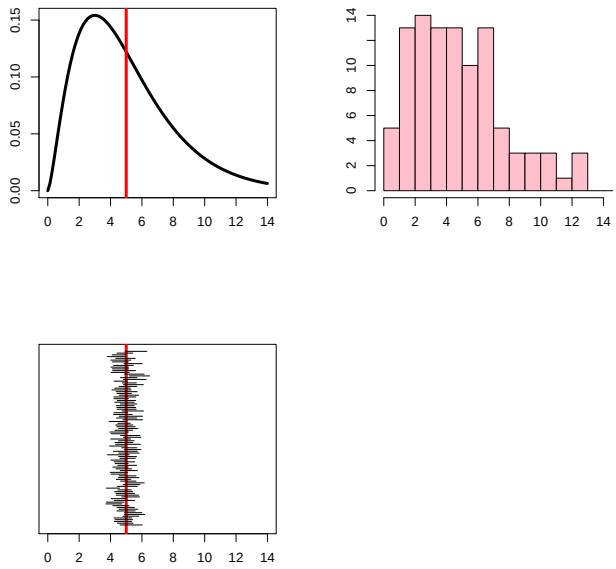

Figure 2: Top left: density of a χ^2 with 5 degrees of freedom. The vertical line shows the mean. Top right: $n = 100$ draw from the distribution. Bottom left: Confidence intervals from simulation.

replacement from $X_1, \ldots, X_n$. *The new sample is denoted by* $X_1^*, \ldots, X_n^*$. *Compute the sample covariance matrix* S^* *of the new data and let* $\widehat{\theta}^*$ *denote the largest eigenvector of* S^*. *Repeat this process* B *times where* B *is typically about 10,000. This yields bootstrap values* $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$. *The standard deviation of* $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ *is an estimate of the standard error of the original estimator* $\widehat{\theta}$.

Figure 3 shows a PCA analysis of US arrest data. The last plot shows bootstrap replications of the first principal component.

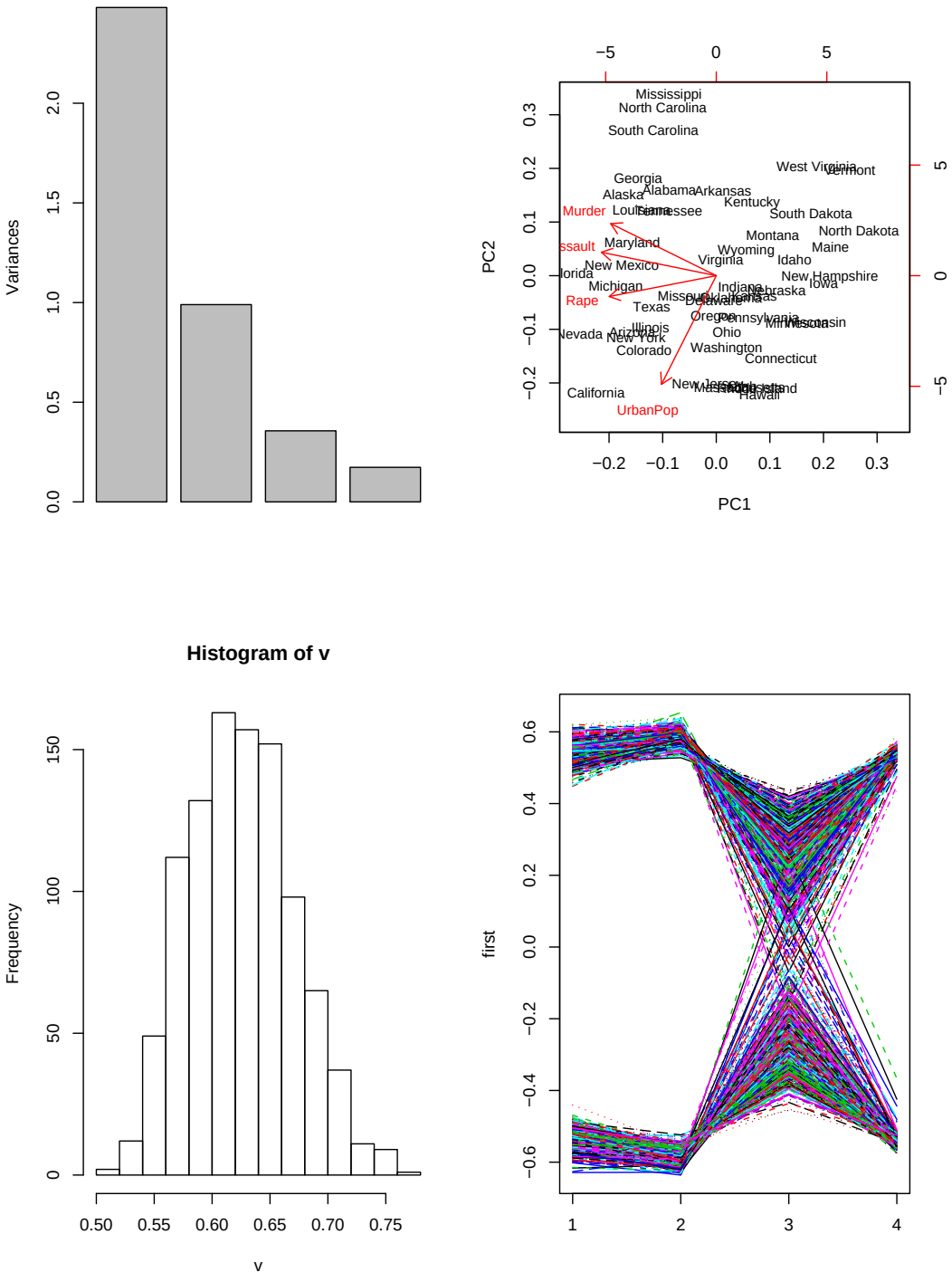

Figure 3: US Arrest Data