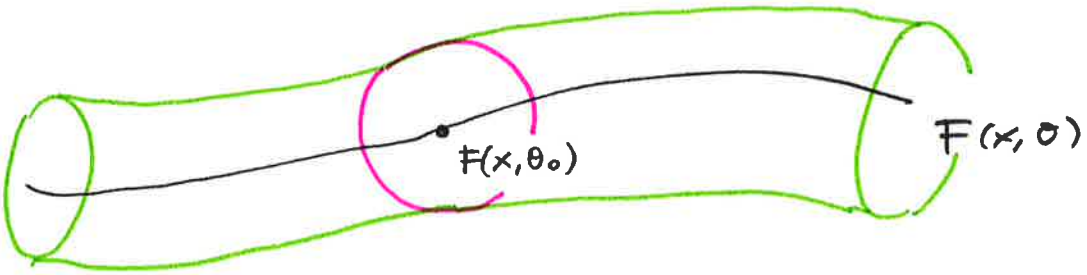
Handout 12 - Parametric Goodness-of-Fit

1 GOF when estimating parameters

In Handouts 11 and 11a, we have seen how to test a distribution, F, which is fully specified, i.e., it does not depend on unknown parameters. However, as you may expect, in many practical scenarios we are actually interested in testing parametric distributions $F(x;\theta)$. For simplicity, here we will consider the case where X and θ are one-dimensional. Extensions to the case where X is multidimensional are trivial (based on the material covered in Handouts 11 and 11a), and extensions to the case where θ is multidimensional are also possible (and not difficult) once one understands the concepts outlined in this Handout.

But what is the difference between testing $F(x;\theta)$, when θ is fixed to a specific (known) value, and testing $F(x;\theta)$ when θ is unknown? Let's look at this from a geometrical perspective:



From the picture above it is clear that if θ is fixed to some specific value θ_0, our null hypothesis consists of one specific points on the manifold above. This means that we are basically testing from deviations from this point within a ball centered at such point. Conversely, if θ is unknown, our null hypothesis consists of the entire manifold and we test from deviations from it within a tube around it.

As one may expect, to test

$$H_0 : X \sim F(x;\theta) \quad \text{versus} \quad H_0 : X \not\sim F(x;\theta).$$

we can rely on the (parametric) empirical process

$$v_n(x,\widehat{\theta};F) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbb{1}_{\{x_i\leq x\}} - F(x;\widehat{\theta})], \tag{1}$$

The test is then performed by taking functionals of it and simulating its distribution using the parametric bootstrap.

Notice that, since the estimation of $\widehat{\theta}$ introduces additional variability, even if X is one-dimensional, we cannot recover distribution-freeness by applying the probability integral transform. That is because,

depends on the estimator of θ

$$\hat{u} = F(x, \hat{\theta}) \not\sim \text{Uniform}[0,1]$$

The natural question arising at this point is then: how can we perform distribution-free parametric goodness-of-fit tests? Is that even possible? The answer is yes and, once again, we can exploit a slight modification of the K-2 transform we have seen in Handout 11a.

To construct distribution-free parametric tests we begin by rewriting the process in equation (1) as a function-parametric empirical process, that is, an empirical process indexed by functions which depend on unknown parameters. Specifically, we define

$$\hat{\phi}_x(x_i) = \phi_x(x_i; \hat{\theta}) = \left[\mathbb{1}_{\{x_i \le x\}} - F(x; \hat{\theta})\right]$$

where $\widehat{\theta}$ is taken to be the MLE of θ, and we rewrite the process in (1) as

$$V_n(\hat{\phi}_x, F) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\phi}_x(x_i)$$

Now let's expand this process using Taylor's expansion around the true value of θ. Specifically, let

$$\phi_x(x_i; \theta) = [\mathbb{1}_{\{x_1 \le x\}} - F(x; \theta)]$$

and notice that

$$\frac{d}{d\theta}\widehat{\phi}_x(x_i;\theta)|_{\widehat{\theta}=\theta} = \frac{d}{d\theta}\phi_x(x_i;\theta) = -\frac{d}{d\theta}F(x,\theta) = -\frac{d}{d\theta}\int_{-\infty}^{x} f(t,\theta)\,dt$$

$$= -\int_{-\infty}^{x} \underbrace{\frac{d}{d\theta} f(t,\theta)}_{\dot{f}(t,\theta)}\,dt = -\int_{-\infty}^{x} \dot{f}(t,\theta)\,dt$$

$$= \boxed{-\int_{-\infty}^{x} \frac{\dot{f}(t,\theta)}{f(t,\theta)}\,dF(t,\theta)}$$

It follows that

$$v_n(\widehat{\phi}_x; F) = v_n(\phi_x; F) + (\widehat{\theta} - \theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d}{d\theta}\widehat{\phi}_x(x_i;\theta)|_{\widehat{\theta}=\theta} + error \tag{2}$$

$$= v_n(\phi_x; F) - (\widehat{\theta} - \theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_{-\infty}^{x}\frac{\dot{f}(t;\theta)}{f(t;\theta)}dF(t;\theta) + error \tag{3}$$

$$= v_n(\phi_x; F) - (\widehat{\theta} - \theta)\sqrt{n}\int_{-\infty}^{x}\frac{\dot{f}(t;\theta)}{f(t;\theta)}dF(t;\theta) + error \tag{4}$$

Now we consider the asymptotic expansion of $(\widehat{\theta} - \theta)$.

- **What would that be?** (recall $\widehat{\theta}$ is the MLE).

$$\sqrt{n}(\hat{\theta} - \theta) \sim \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\theta^{-1}\frac{d}{d\theta}\log f(x_i;\theta)$$

↑ Fisher information

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\theta^{-1}\frac{\dot{f}(x_i;\theta)}{f(x_i;\theta)}$$

$$I_\theta^{-1/2}\, a(x_i;\theta)$$

Let $a(x;\theta)$ be the normalized scored function, i.e.,

$$a(x;\theta) = \Gamma_\theta^{-1/2}\frac{\dot{f}(x;\theta)}{f(x;\theta)}$$

we can now rewrite the process in (4) as

$$v_n(\widehat{\phi}_x; F) = v_n(\phi_x; F) - \frac{1}{n}\sum_{i=1}^{n}\Gamma_\theta^{-1}\frac{\dot{f}(x_i;\theta)}{f(x_i;\theta)}\sqrt{n}\int_{-\infty}^{x}\frac{\dot{f}(t;\theta)}{f(t;\theta)}dF(t;\theta) + error \tag{5}$$

(6)

$$= v_n(\phi_x; F) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Gamma_\theta^{-1/2}\frac{\dot{f}(x_i;\theta)}{f(x_i;\theta)}\int_{-\infty}^{x}\Gamma_\theta^{-1/2}\frac{\dot{f}(t;\theta)}{f(t;\theta)}dF(t;\theta) + error \tag{7}$$

$$= v_n(\phi_x; F) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(x_i;\theta)\int_{-\infty}^{x} a(t;\theta)dF(t;\theta) + error \tag{8}$$

Finally, I claim that

$$< \phi_x, a >_F = \int_{\infty}^{x} a(t;\theta)dF(t;\theta)$$

(you may prove it as an exercise).

It follows that

$$V_n(\hat{\phi}_x, F) = V_n(\phi_x, F) - V_n(a, F)\langle \phi_x, a\rangle_F + \text{error} \tag{9}$$

$$V_n(a, F) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(x_i, \theta)$$

$$\sim \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_x(x_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(x_i;\theta)\langle \phi_x, a\rangle_F$$

$\hookrightarrow \mathbb{1}_{\{x_i \le x\}} - F(x;\theta)$

where one can show that the error is $o_p(1)$. The limiting process is therefore a projection of $v(\phi_x; F)$ parallel to the score function a. i.e., a projected Brownian motion (and thus it is still a Gaussian process).

- **From a computational perspective, can you think of one advantage of relying on (9), instead of (1)?**

The RHS doesn't depend on $\hat{\theta}$
$\Rightarrow$ when doing the bootstrap we don't need to re-estimate the parameters and we don't need to re-evaluate $F(x, \hat{\theta})$

We can also re-write the process in equation (9) (minus the error) as

$$v_n(\tilde{\phi}_x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\phi}_x(x_i;\theta) \tag{10}$$

with

$$\tilde{\phi}_x(x_i;\theta) = \phi_x(x_i;\theta) - a(x_i;\theta)\langle a, \phi_x\rangle_F$$

and thus the process in equation (10) is an empirical process indexed by functions $\tilde{\phi}_x \in \mathcal{L}_\perp(F) \subset \mathcal{L}(F) \subset L^2(F)$. Where $\mathcal{L}_\perp(F)$ is the set of functions which are square integrable w.r.t. F, are orthogonal to one (have mean zero) under F, and are orthogonal to a w.r.t. F, i.e.,

$\langle \tilde{\phi}_x, \tilde{\phi}_x\rangle_F < +\infty$ $\quad \langle \tilde{\phi}_x, 1\rangle_F = 0$ $\quad \langle \tilde{\phi}_x, a\rangle_F = 0$

$$\langle \tilde{\phi}_x, a\rangle_F = \langle \phi_x - a\langle a, \phi_x\rangle_F, a\rangle_F$$

$$= \langle \phi_x, a\rangle_F - \underbrace{\langle a, a\rangle_F}_{=1}\langle a, \phi_x\rangle_F$$

$$= 0$$

2 Towards distribution-freeness

All the consideration above hold for any arbitrary distribution function. To perform distribution-free goodness-of-fit we will consider a *reference distribution* Q, and the function-parametric empirical process:

$$Q(x;\beta)$$

$$V_n(\tilde{\psi}_x;\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_x(x_i,\beta)$$

$$\tilde{\psi}_x(x_i,\beta) = \psi_x(x_i,\beta) - b(x_i,\beta)\langle b, \psi_x\rangle_Q$$

$$\psi_x(x_i;\beta) = \left[\mathbb{1}_{\{x_i\le x\}} - Q(x,\beta)\right]$$

$$b(x_i,\beta) = I_\beta^{-1/2}\frac{d}{d\beta}\log q(x;\beta)$$

$$\tilde{\psi}_x \in \mathcal{L}_\perp(Q) \quad b_j \in \mathcal{L}_\perp(Q)$$

Now the question is: how do we construct an empirical process which, under F, has the same limiting distribution as this process? Well, all we need to do is to re-adapt the K-2 transform. Specifically, we proceed according to the following steps:

Step 1 - Map $\tilde{\psi}_x$ and b into $L^2(F)$ via the isometry

$$l(t;\theta,\beta) = \sqrt{\frac{q(t;\beta)}{f(t;\theta)}},$$

where q and f are the densities of Q and F, respectively. Obtain $l\tilde{\psi}_x$ and lb, both $\in L^2(F)$.

Step 2 - Map $l\tilde{\psi}_x$ and lb into $\mathcal{L}(F)$ by means of the the unitary operator

$$K = I - \frac{1-l}{1-\langle l,1\rangle_F}\langle 1-l,\cdot\rangle_F, \tag{11}$$

and obtain $Kl\tilde{\psi}_x$ and $c = Klb$, both $\in \mathcal{L}(F)$.

Step 3 - Map c into a and $Kl\tilde{\psi}_x$ into functions in $\mathcal{L}_\perp(F)$ by means of the the unitary operator

$$U_{a,c} = I - \frac{a-c}{1-\langle a,c\rangle_F}\langle a-c,\cdot\rangle_F. \tag{12}$$

Notice that $U_{a,c}$ maps a into c, c into a, and leaves functions orthogonal to a and c unchanged. Obtain, $UKl\tilde{\psi}_x \in \mathcal{L}_\perp(F)$. To see that:

are the $Uke\tilde{\psi} \in \mathcal{L}_1(F)$?

$$\begin{aligned}
\langle Uke\tilde{\psi}, a\rangle_F &= \langle Uke\tilde{\psi}, Uc\rangle_F \\
&= \langle ke\tilde{\psi}, c\rangle_F \\
&= \langle ke\tilde{\psi}, keb\rangle_F \\
&= \langle e\tilde{\psi}, eb\rangle_F \\
&= \int e^2(t)\,\tilde{\psi}_x(t)\,b(t,\beta)\,dF(t) \\
&= \int q(t,\beta)\,\tilde{\psi}_x(t)\,b(t,\beta)\,dt = \langle \tilde{\psi}_x, b\rangle_Q = 0
\end{aligned}$$

Similarly to what we have seen in Handout 11a, it follows that $v_n(UKl\tilde{\psi}_x; F)$ and $v_n(\tilde{\psi}_x; Q)$, converge to a Gaussian process $v(\tilde{\psi}_x; Q)$ with mean and covariance:

$$\langle Uke\tilde{\psi}, 1\rangle_F = \langle \tilde{\psi}, 1\rangle_Q = 0$$

$$\langle Uke\tilde{\psi}_x, Uke\tilde{\psi}_{x'}\rangle_F = \langle \tilde{\psi}_x, \tilde{\psi}_{x'}\rangle_Q < +\infty$$

Hence we construct an entire family of test statistics for testing $H_0 : X \sim F(x;\theta)$ versus $H_0 : X \not\sim F(x;\theta)$, i.e.,

- **Kolmogorov's statistics:** $\sup|v_n(UKl\tilde{\psi}_x; F)| \xrightarrow{d} \sup|v(\tilde{\psi}_x; Q)|$.
- **Cramer von Mises statistics:** $\int |v_n(UKl\tilde{\psi}_x; F)|^2 dQ(x) \xrightarrow{d} \int |v(\tilde{\psi}_x; Q)|^2 dQ(x)$.
- **Anderson-Darling statistics:** $\int |\frac{v_n(UKl\tilde{\psi}_x;F)}{\sqrt{Q(x)(1-Q(x))}}|^2 dQ(x) \xrightarrow{d} \int |\frac{v(\tilde{\psi}_x;Q)}{\sqrt{Q(x)(1-Q(x))}}|^2 dQ(x)$.

Once again, all we have to do to perform the test is to simulate the distribution of the functionals of $v_n(\psi_x; Q)$ under Q and use it to assign significance to the values of the functionals of $v_n(UKl\psi_x; F)$ observed on the data.

For students enrolled in Fall 2022:

- Please note that the policies listed below will be similar to those of your exam on October 26, 2022. The only difference is that, for your first midterm, you are only allowed a one page front only cheat sheet
- The questions relevant to you to prepare for your first midterm have been highlighted in yellow throughout this exam. The remaining questions refer to material that is not going to be covered in your first midterm (but will be in the second midterm).

Written Exam

December 3, 2021

Do not open the exam until you are instructed to do so.

Exam policy

- Write your name and your University ID below.
- On your desk you are only allowed the following items:
  - up to three black or blue pens,
  - one 8.26" by 11.69" page cheat sheet (front and back),
  - a bottle of water.
- The exam will last 50 minutes.
- You are not allowed to leave the room within the first 15 minutes from the beginning of the exam and within the last 10 minutes.
- If you have completed the exam before the last 10 minutes, close your exam, stay seated and rise your hand.

Name: **ID:**

Q1. List five statistical learning tools (among those we have seen in the course) which allow us to perform classification when the decision boundaries are expected to be non-linear.

(a) Classification trees
(b) SVM
(c) KNN
(d) QDA
(e) GAM

Q2. Suppose you are given a sample of n i.i.d. observations from a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ with independent components, and assume $p < n$. When estimating the covariance matrix of $\boldsymbol{X}$ using the sample covariance matrix, $\boldsymbol{S}$, which is the problem arising as p/n increases? Which estimator would you recommend to overcome this problem?

2 points

As the ratio p/n increases, S becomes more and more unstable as an estimator of true covariance matrix

This problem can be addressed by relying on a shrinkage estimator such as Ledoit-Wolf of the form $\hat{\Sigma} = \alpha S + (1-\alpha) I$

Q3. If we were to rewrite the support vector classifier as a special case of support vector machines, what would be the respective kernel? (Please note that you are not asked to derive the kernel, simply state what that is). Would the kernel be a reproducing kernel? Justify your answer.

The respective kernel would be the linear kernel

$$K(\underline{x}_i, \underline{x}_j) = \underline{x}_i' \underline{x}_j$$

The Gram matrix associated with this kernel is

$$\underset{s \times p}{X} \underset{p \times s}{X'} \qquad s = \#\text{ of support vectors}$$

which is non-negative

$$\forall a \in \mathbb{R}^s \quad a' X X' a = \underbrace{(X'a)'}_{b} (X'a) = b'b \geq 0$$

$\Rightarrow K(\underline{x}_i, \underline{x}_j)$ is a Mercer kernel

$\Rightarrow$ it has the reproducing property

Q4. Consider the model $Y = \boldsymbol{X}\beta + \epsilon$ where the predictors in $\boldsymbol{X}$ are linearly independent. Write down the step necessary to estimate its mean square error using the "bad" bootstrap. Please make sure you define all the quantities involved. Also, explain what is a drawback of this procedure and how we could modify it in order to address the problem.

6 points

4 points

Step 1) Draw B bootstrap datasets

$$Z^{(b)} = (Y_1^{(b)}, \underline{X}_1^{\prime(b)}) \ldots\ldots (Y_n, \underline{X}_n^{\prime(b)})$$

with replacement from the original dataset

Step 2) We fit $Y = X\beta + \varepsilon$ on each $Z^{(b)}$ and obtain $\hat{\beta}^{(b)}$

Step 3) Predict each observation in the original dataset via

$$\hat{Y}_i^{(b)} = \underline{x}_i^T \hat{\beta}^{(b)}$$

Step 4) Obtain the bootstrap estimate of the test MSE

$$\widehat{MSE} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(b)})^2$$

Drawback: Training and validation sets all contain the same information. This will lead to overly optimistic estimates of the prediction error

1 point

Possible solution: Instead of predicting the observations in the original dataset, in Step 3, we predict the observations that were "left out" from the sampling procedure and we use those to compute $\widehat{MSE}$ in Step 4

1 point

Q5. What are the two main advantages of using KDE instead of histograms to estimate probability density functions?

(1) it provides a continuous estimate of the pdf

(2) it provides a local estimate

Q6. Suppose we are aiming to predict $Y \in \{-1, 1\}$ using the information carried by a continuous predictor X. Write down the main steps of AdaBoost, when the weak classifiers G_b consist of simple logistic regression models. Please make sure you define all the quantities involved.

Step 1: Initialise the observation weights $w_i = \frac{1}{n}$

Step 2: For $b = 1, \ldots, B$

(a) Fit $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \quad \forall i = 1, \ldots, n$

by maximizing the weighted log-likelihood with weights w_i and we obtain $\hat{\pi}_i$

(b) $G_b(x_i) = 2 \mathbb{1}_{\{\hat{\pi}_i \geq 0.5\}} - 1$

(c) Compute $\overline{err}_b = \frac{\sum_{i=1}^{n} w_i \mathbb{1}_{\{y_i \neq G_b(x_i)\}}}{\sum w_i}$

(d) Compute $\alpha_b = \log\left(\frac{1 - \overline{err}_b}{\overline{err}_b}\right)$

(e) Set $w_i = w_i \exp\{\alpha_b I_{\{y_i \neq G_b(x_i)\}}\}$

Step 3: Output

$$G(x) = \text{sign}\left[\sum_{b=1}^{B} \alpha_b G_b(x)\right]$$