Summary of Ingster's and Suslina's Book

Akshay Krishnamurthy
akshaykr@cs.cmu.edu

February 18, 2014

Overview

This document will present a summary of results and key ideas from the book **Nonparametric Goodness-of-Fit Testing under Gaussian Models** by Ingster and Suslina. The presentation will be quite concise – refer to the book itself for details.

1 Chapter 1

Chapter 1 is an overview of the problems of study. Most of this material is restated in chapter 2 where it is presented in some more detail.

2 Chapter 2

The main problem of study is the goodness of fit testing problem in the gaussian sequence model. Here we are given a realization of a gaussian random vector:

$$X = v + \eta, \quad v \in \mathbb{R}^n \quad \text{or } v \in \ell^2$$

where η is a a sequence of standard gaussian random variables and v is the mean. $\ell^2 = \{v : \sum_i v_i^2 < \infty\}$ is the usual Hilbert space of square summable sequences.

The equivalence between the gaussian sequence model and the functional model is established. Here:

$$dX_\epsilon(t) = s(t)dt + \epsilon dW(t), \;\; s \in L_2(0,1)$$

$W(t)$ is a realization of the standard Weiner process and $s(t)$ is the unknown signal. The equivalence stems from the Fourier expansion of X.

In goodness of fit testing, we are interested in testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$ where Θ_0 is a subset of some parameter space Θ. Usually we are interested in $\Theta_0 = \{\theta\}$ is a point mass. We will also denote $\mathscr{P}_0, \mathscr{P}_1$ the set of distributions in the range of the parameterization Θ_0, Θ_1, respectively.

A test is a measurable function (possible randomized) ψ taking values in $[0,1]$ where $\psi(X)$ denotes the probability of rejecting the null given that the observation was X. The standard terminology is introduced (Type I, Type II errors, simple/composite hypothesis etc.)

Two ways to compare tests based on the Type I and Type II errors are introduced. We need several definitions:

$$\begin{aligned} \alpha(\psi,P) &= \mathbb{E}_P\psi, P\in\mathscr{P}_0 && (1)\\ \beta(\psi,P) &= \mathbb{E}_P 1-\psi, P\in\mathscr{P}_1 && (2)\\ \gamma_t(P_0,P_1) &= \inf_\psi \gamma_t(\psi;P_0,P_1) = \inf_\psi t\alpha(\psi,P_0)+\beta(\psi,P_1) && (3)\end{aligned}$$

$\gamma_t(\psi)$ defines the **risk** of the estimator ψ with simple null and alternative. One way to compare tests is to look at their risk.

The other way is due to Neyman and Pearson. The idea is to compare the Type II error rates β over all tests with Type I error rate bounded by α. Defining the set $\Psi_\alpha(P_0)=\{\psi:\alpha(\psi,P_0)\le\alpha\}$ we can look at:

$$\beta(\alpha)=\beta(\alpha,P_0,P_1)=\inf_{\psi\in\Psi_\alpha(P_0)}\beta(\psi,P_1)$$

Again a test $\psi_\alpha\in\Psi_\alpha$ is optimal here if it provides the infimum.

2.1 Neyman-Pearson Lemma

The Neyman-Pearson lemma determines the structure of the optimal test when both the null and alternative are simple. Suppose that P_0 has density f_0 and P_1 has density f_1 (with respect to say the Lebesgue measure). For a parameter t define:

$$\mathscr{X}_t^- = \{X\in\mathscr{X}: f_1(X)<tf_0(X)\}, \mathscr{X}_t^+ = \{X\in\mathscr{X}: f_1(x)>tf_0(x)\}$$

We have to be mindful of the set where $f_1(X)=tf_0(X)$, refer to the text for details.

Lemma 1. *(Neyman-Pearson) The optimal tests, under γ_t, $\psi^{(t)}$ are of the form:*

$$\psi^{(t)}(X)=\begin{cases}0 \text{ if } & X\in\mathscr{X}_t^-\\ 1 \text{ if } & X\in\mathscr{X}_t^+\end{cases} \tag{4}$$

Essentially the same thing is true for the optimal test ψ_α under $\beta(\alpha)$ with $t=t_\alpha$

The key relation in the proof is the connection to the total variational distance. Recall that the total variation distance (when P_0,P_1 are dominated by P):

$$||P_1-P_0||_1=\int_{\mathscr{X}}|f_1(x)-f_0(x)|P(dx)$$

The proof shows that:

$$\gamma_1(P_0,P_1)=1-\frac{1}{2}||P_1-P_0||_1$$

which will be essential in constructing lower bounds.

It should be noted that the functions γ,β are continuous with respect to TV distance (Proposition 2.1). This result plays some role later.

2.2 Bayesion vs Minimax view

Even as it is, when the null and/or alternative are composite, the functions γ, β are not sufficient to compare tests. In the bayesian view, we fix a prior π_0 on the space $\mathscr{P}_0$ and a prior π_1 on the space $\mathscr{P}_1$. Now we can redefine the quantities γ, α, β with respect to the priors:

$$\begin{aligned}
\alpha(\psi, \pi_0) &= \mathbb{E}_{P\sim\pi_0}\alpha(\psi, P) \\
\beta(\psi, \pi_1) &= \mathbb{E}_{P\sim\pi_1}\beta(\psi, P) \\
\gamma(t; \pi_0, \pi_1) &= \inf_{\psi\in\Psi} \gamma_t(\psi; \pi_0, \pi_1) = \inf_{\psi\in\Psi} t\alpha(\psi, \pi_0) + \beta(\psi, \pi_1) \\
\beta(\alpha; \pi_0, \pi_1) &= \inf_{\psi\in\Psi_{\alpha,\pi_0}} \beta(\psi, \pi_1)
\end{aligned}$$

If we define the mixtures $P_{\pi_0}(A) = \mathbb{E}_{P\in\pi_0}P(A), A \in \mathscr{A}$ and P_{π_1} similarly, then we are back in the simple vs. simple setting and can appeal to the Neyman Pearson Lemma. In particular:

$$\gamma(1) = 1 - \frac{1}{2}||P_{\pi_0} - P_{\pi_1}||_1 = 1 - \frac{1}{2}\mathbb{E}_{P_{\pi_0}}|L - 1|$$

where $L = dP_{\pi_1}/dP_{\pi_0}$ is the likelihood ratio statistic in the Bayesian settup.

In the finite-dimensional gaussian model, we can prove that the optimal test $\psi^{(t)}, \psi_\alpha$ are non-randomized tests of the form $\mathbf{1}_{\mathscr{X}}$ for some measurable set $\mathscr{X} \subset \mathbb{R}^n$ (Proposition 2.4). The strategy is to show that the LRT is an analytic function which gives us all sorts of nice properties. Unfortunately in many cases we cannot work with the bayesian likelihood ratio test.

One example when we can is when π_0 is a point mass and π_1 is uniformly distributed over the sphere $S^{n-1}(\rho) = \{v \in \mathbb{R}^n : ||v|| = \rho\}$. In this case the optimal test depends on the norm $||X||$ and the performance is governed by a the distribution function of ξ_n^2 random variables.

In the minimax approach, we replace α, β, γ with their worst-case equivalents:

$$\begin{aligned}
\alpha(\psi, \mathscr{P}_0) &= \sup_{P\in\mathscr{P}_0} \alpha(\psi, P) \\
\beta(\psi, \mathscr{P}_1) &= \sup_{p\in\mathscr{P}_1} \beta(\psi, P) \\
\gamma_t(\mathscr{P}_0, \mathscr{P}_1) &= \inf_{\psi\in\Psi} \gamma_t(\psi) = \inf_{\psi\in\Psi} t\alpha(\psi) + \beta(\psi) \\
\beta(\alpha, \mathscr{P}_0, \mathscr{P}_1) &= \inf_{\psi\in\Psi_\alpha} \beta(\psi)
\end{aligned}$$

In order to establish the connection with the total variation distance we need to establish some geometric properties on α, γ, β. These functions are linear continuous functions in both ψ and P_0, P_1. The functions $\alpha(\psi, \mathscr{P}_0), \beta(\psi, \mathscr{P}_1), \gamma_t(\psi, \mathscr{P}_0, \mathscr{P}_1)$ are all convex in ψ. And $\beta(\alpha)$ is convex nonincreasing continuous in α and $\gamma(t)$ is concave nondecreasing continuous in t. Moreover, $\beta(\alpha)$ and $\gamma(t)$ look something like conjugates of each other.

$$\gamma(t) = \inf_{\alpha\in[0,1]} t\alpha + \beta(\alpha), \quad \beta(\alpha) = \sup_{t\in[0,\infty)} (\gamma(t) - t\alpha)$$

The main result of this section is the connection to Bayesian testing.

Theorem 2 (Theorem 2.1)**.** *Define* $[A]$ *to be the convex hull of* A*. Then under some conditions*

$$\gamma(t) = \sup\{\gamma(t, P_0, P_1) : P_0 \in [\mathscr{P}_0], P_1 \in [\mathscr{P}_1]\} \quad (5)$$

$$\beta(\alpha) = \sup\{\beta(\alpha, P_0, P_1) : P_0 \in [\mathscr{P}_0], P_1 \in [\mathscr{P}_1]\} \quad (6)$$

$$\gamma(1) = 1 - \frac{1}{2}|[\mathscr{P}_0], [\mathscr{P}_1]|_1 \quad (7)$$

Moreover, as long as $\mathscr{P}_0, \mathscr{P}_1$ *are compact in TV, then there exists priors* π_0, π_1 *such that the minimax tests are optimal bayesian tests. These priors are the* ***least favorable priors****.*

Essential in proving the theorem is the saddle-point characterization of concave-convex functions:

$$\min_{x \in X} \sup_{y \in Y} f(x, y) = \sup_{y \in Y} \min_{x \in X} f(x, y)$$

when X, Y are convex, $f(x, y)$ is convex in x and concave in y, X is compact and f is semi-continuous in x for each y.

Notice that already we have a strategy for constructing lower bounds since:

$$\gamma \geq 1 - \frac{1}{2}|P_{\pi_0} - P_{\pi_1}|_1$$

for any priors π_0, π_1 that concentrate on $\mathscr{P}_0, \mathscr{P}_1$ respectively.

Typically constructing least favorable priors is challenging. However we can sometimes certify that a prior is least favorable.

Theorem 3 (Theorem 2.3)**.** *Priors* π_0, π_1 *are least favorable when* $\pi_0(S_0) = \pi_1(S_1) = 1$ *for measurable sets* $S_0 \subset \mathscr{P}_0, S_1 \subset \mathscr{P}_1$ *and the bayesian test* $\psi^{(t)}, \psi_\alpha$ *satisfy:*

$$\gamma_t(\psi^{(t)}; \mathscr{P}_0, \mathscr{P}_1) = \gamma_t(\psi^{(t)}; P_0, P_1), \alpha(\psi_\alpha, \mathscr{P}_0) = \alpha(\psi_\alpha, P_0), \beta(\psi_\alpha, \mathscr{P}_1) = \beta(\psi_\alpha, P_1)$$

For all $P_0 \in S_0, P_1 \in S_1$*. If this is true then* $\psi_\alpha, \psi^{(t)}$ *are minimax.*

Essentially, if we can construct a prior so that the bayesian test performs uniformly minimax over the prior then we have found the least favorable prior.

Example 1 (Convex Alternative)**.** *In the finite dimensional gaussian model if the null is* $H_0 : v = 0$ *and the alternative corresponds to means lying in some convex set. Then the minimax rate is governed by the minimum length point* $v^\star = \inf_{v \in V} ||v||$*. The least favorable prior is a dirac mass* $\delta_{v^\star}$*. This can be easily extended to composite null hypothesis.*

The concrete example is when the alternative V *is the exterior of the positive part of an* ℓ_p *ball (which we we call* $\check{D}^+_{n,p}(\rho_n)$*). I.e.* $v \in \mathbb{R}^n_+, \left(\sum_{i=1}^n v_i^p\right)^{1/p} \geq \rho_n, \rho_n > 0, p \in (0, 1]$ *When* $p \in (0, 1]$ *the alternative set is convex and it is easy to check that the minimax rate of testing is* $\rho_n^\star = n^{1/p - 1/2}$*. In other words* $\gamma(V_n) \to 0$ *iff* $\rho_n / \rho_n^\star \to \infty$ *and* $\gamma(V_n) \to 1$ *iff* $\rho_n / \rho_n^\star \to 0$*.*

If we try to extend this to the infinite dimensional setting we find that the point 0 *is actually in the closure of the alternative and that it is impossible to distinguish the null and alternative. This is a key observation echoed throughout the chapter – it is necessary to consider smaller alternatives if we want to obtain non-trivial results in the non-parametric setting.*

2.3 Asymptotics

Usually we'll only care about asymptotics. There will be an asymptotic parameter $\epsilon \to \epsilon_0$ and we are interested in the performance of the testing problem as a function of ϵ. The setup contains a sequence Θ_ϵ of parameter spaces with nulls $\Theta_{\epsilon,0}$ and alternatives $\Theta_{\epsilon,1}$ and we would like to construct a sequence of tests $\{\psi_\epsilon\}$ with good asymptotic performance. Specifically we want $\gamma_{\epsilon,t}(\psi_\epsilon^{(t)}) \to \gamma_\epsilon(t) + o(1)$ as $\epsilon \to \epsilon_0$. The goal is similarly augmented in the Neyman-Pearson seting.

It is enough to consider sequences of priors $\pi_{\epsilon,0}, \pi_{\epsilon,1}$ that asymptotically put all of their mass on $\Theta_{\epsilon,0}, \Theta_{\epsilon,1}$. Again we have the lower bound:

$$\gamma_\epsilon \geq 1 - \frac{1}{2}|P_{\pi_{\epsilon,0}} - P_{\pi_{\epsilon,1}}| + o(1)$$

It is also the case that instead of total variation distance, we can work with the L_2 metric.

Proposition 4 (Proposition 2.12)**.** *If* $||P_{\epsilon,1} - P_{\epsilon,0}||_2 \to 0$*, then* $||P_{\epsilon,1} - P_{\epsilon,0}||_1 \to 0$ *so* $\gamma \to 0$*. If* $||P_{\epsilon,1} - P_{\epsilon,0}||_1 = O(1)$ *then* $\limsup ||P_{\epsilon,1} - P_{\epsilon,0}||_2 < 2$ *and* $\liminf \gamma > 0$*.*

So now we can establish lower bounds by finding priors π_0, π_1 whose mixtures are close in L_2.

2.4 Minimaxity in Goodness of Fit Testing

This section presents several examples. In goodness of fit testing we are typically concerned with the following problem:

$$H_0 : \theta \in \Theta_{\epsilon,0} = \{\theta_0\} \;\; H_1 : \theta \in \Theta_{\epsilon,1} = \{\theta : h_1(\theta, \Theta_{\epsilon,0}) \geq r_\epsilon\}$$

Where h_1 is some notion of distance. We want to know: What is the radius r_ϵ under which we can obtain minimax distinguishability, $\gamma_\epsilon \to 0$?

Example 2 (L_p norm on distribution function)**.** *Consider the i.i.d. sample model with density* f *on the interval* $[0,1]$ *with the following hypothesis:*

$$H_0 : f = f_0 : f_0(x) = 1, H_1 : f \in \mathscr{F}_{r_N,p} = \{f : ||F_{f_0} - F_{f_1}||_p \geq r_N\}$$

where F_f *is the cumulative distribution function for the density* f*. Here we are testing the uniform distribution from the set of distributions that are far away (in terms of CDF), given* N *samples. These are* ***Kolmogorov norms*** *and result in Kolmogorov / Cramer-von-Mises-Smirnov tests.*

Here the rates are ***classical*** *in that if* $N^{1/2} r_n \to \infty$ *then tests based on the empirical distribution function are minimax consistent and if* $N^{1/2} r_n = O(1)$ *then* $\liminf_{N\to\infty} \gamma_N > 0$*. (Proposition 2.14)*

Example 3 (Non-classical problems)**.** *In the n-dimensional gaussian model, consider testing the simple hypothesis* $H_0 : v = 0$ *against the alternative* $H_1 : v \in \breve{D}^n(\rho_n)$ *(the exterior of a sphere of radius* ρ_n*). Here the distinguishability conditions are of the form* $\rho_n^\star = n^{1/4}$*. This problem is classical.*

However in the Gaussian sequence model, testing $H_0 : v = 0$ from $H_1 : v \in \check{D}_{(\ell)}(\rho) = \{v \in \ell^2 : ||v|| \geq \rho\}$, then $\gamma = 1$ for any $\rho > 0$. The same is true in the functional gaussian model and for any norm $0 < p \leq \infty$ which parameterizes the alternative. Similarly in the i.i.d. sample model testing $H_0 : f_0(t) = 1$ versus $H_1 : f \in \check{D}_{p,f}(r, f_0) = \{f : ||f - f_0||_p \geq r\}$, we have $\gamma = 1$ for any $r \in (0, 1)$.

Observe the distinction in the i.i.d. sample setting – if the alternative constraint is placed on the "-1st" derivative of the density (the CDF) then we get classical rates. However if the norm is based on the density itself, then the set is somehow too wide and we get triviality. To make progress, we need to place additional constraints on Θ_1, which usually are norm constraints $h_2(\theta) \leq R$ for some other norm h_2.

2.5 Norms

As mentioned, we will need to place some constraints on the alternative space Θ_1 and this is usually done in the form of "norm" constraints. In this section we define several norms used throughout the book.

Definition 5. *Let $\bar{a} = \{a_i, i \in I\}$ be a given sequence. The $\ell_{\bar{a},p}$ norm is:*

$$||v||_{\bar{a},p} = \begin{cases} \left(\sum_i |v_i a_i|^p \right)^{1/p} & \text{if } 0 < p < \infty \\ \sup_i |v_i a_i| & \text{if } p = \infty \end{cases} \tag{8}$$

*The ℓ_p-**ellipsoid** $D_{\bar{a},p}(\rho) = \{v \in \ell^2 : ||v||_{\bar{a},p} \leq \rho\}$ with exterior and surface $\check{D}_{\bar{a},p}(\rho), S_{\bar{a},p}(\rho)$. The special case where $a_i = i^r$ for some r is called the **power norm**.*

We are also interested in **Besov Norms** and **Besov Ellipsoids**, which are connected with wavelet bases. We will write a sequence $v \in \ell^2$ as $\{v_k, 1 \leq k \leq K, v_{ij} j > J_0, 1 \leq i \leq 2^j\}$ for two integers $K \geq 1, J_0 \geq 0$.

Definition 6. *The **Besov Norm** $||\cdot||_{r,p,h}$ is defined as:*

$$||v||_{r,p,h} = \left(\left(\sum_{k=1}^{K} |v_k|^p \right)^{h/p} + \sum_{j=J_0+1}^{\infty} 2^{jrh} \left(\sum_{i=1}^{2^j} |v_{ij}|^p \right)^{h/p} \right)^{1/h}$$

This is like taking the ℓ_h norm across the levels and the ℓ_p norm on each level, but with appropriate reweighting on the later levels.

We also need to define norms on the functional space. The **Sobolev Norms** are:

$$\begin{aligned} ||f||_{m,p}^0 &= ||f^{(m)}||_p = \left(\int |f^{(m)}(x)|^p dx \right) \text{ or } \sup_x |f^{(m)}(x)| \\ ||f||_{m,p} &= ||f||_{m,p}^0 + ||f||_p \end{aligned}$$

Here $f^{(m)}$ is the mth derivative of f. Of course for these norms to be defined we need the mth derivatives to exist.

Defining Besov norms in functional space is quite technical. I recommend reading the book.

2.6 Constraints and Rates

In the estimation problem, we usually define the loss in terms of some norm $||\cdot||_{(1)}$ and the smoothness in terms of another $||\cdot||_{(2)}$. For example, consider $||\cdot||_{(1)} = ||\cdot||_2$ and $||\cdot||_{(2)} = ||\cdot||_{2,\eta}$ is a Sobolev norms in functional space. Then the estimation problem has the smoothness constraint $||f||_{2,\eta} \leq R$ and loss is measured as $l_\epsilon(\theta_\epsilon, \theta) = L(r_\epsilon^{-1}||\theta_\epsilon - \theta||_2)$ for some monotonic function L. The rates of estimation are:

$$r_\epsilon \asymp \epsilon^{2\eta/(2\eta+1)}$$

In the iid sample model they are $r_N \asymp N^{-\eta/(2\eta+1)}$.

Going back to the hypothesis testing setting, let us formalize what exactly are the alternatives we are interested in studying. In the sequence space we are interested in alternatives parameterized by either **power norms** or **Besov norms**:

$$V_\epsilon = \{v \in \ell^2 : ||v||_{r,p} \geq \rho_\epsilon, ||v||_{s,q} \leq R_\epsilon\} \quad (9)$$

$$V_\epsilon = \{v \in \ell^2 : ||v||_{r,p,h} \geq \rho_\epsilon, ||v||_{s,q,t} \leq R_\epsilon\} \quad (10)$$

$$(11)$$

We will translate results from this setting to the functional gaussian setting.

In classical settings, we can use **plug-in** estimators as test statistics and achieve minimax distinguishability. For example histogram estimators are minimax for alteratives based on Kolmogorov norm.

However in nonparametric settings, estimation rates are substantially larger than hypothesis testing rates. These are the problems we will study in the sequel.

3 Minimax Distinguishability

This chapter studies the asymptotic regime and executes the general strategy outlined in chapters 1 and 2. First, we study the asymptotic properties of several test statistics which will be instrumental in establishing upper bounds. Then we look at a reduction and number of prior constructions that establish lower bounds. Finally we use these ideas to establish minimax rates in a variety of settings.

3.1 Test Statistics

The **linear test** $\psi_{r,T} = \mathbf{1}_{t_r > T}$ is based on the linear statistic:

$$t_r = \langle x, r \rangle = \sum_i x_i r_i, r \in \ell^2 ||r||^2 = 1$$

The quantity $\langle X, r\rangle \sim \mathcal{N}(\langle \mathbb{E}[X], r\rangle, 1)$ (compare the characteristic functions) meaning that under the null (when $\mathbb{E}[x] = 0$) $t_r \sim \mathcal{N}(0,1)$. This gives a bound on the Type I and Type II errors:

$$\alpha(\psi_{r,T}) = \mathbb{P}_0[\langle x, r\rangle \geq T] = 1 - \Phi(T) = \Phi(-T) \tag{12}$$

$$\beta(\psi_{r,T}, V) = \sup_{v \in V} \Phi(T - \langle r, v\rangle) = \Phi(T - h(V)) \tag{13}$$

where $h(V) = \inf_{v \in V} \langle r, v\rangle$. The quantity $h(V)$ controls the performance of these linear tests. In particular, if $h(V_\epsilon) \to \infty$ then $\beta_\epsilon(\alpha) \to 0$ leading to an asymptotically distinguishing test (Corollary 3.2). Recall that ϵ is the asymptotic parameter.

The χ^2**-test** $\psi_{\epsilon,T_\epsilon}$ is based on the χ^2 statistic and requires a family of non-negative weights $\{w_{\epsilon,i}\}$:

$$\psi_{\epsilon,T_\epsilon} = \mathbf{1}_{t_\epsilon > T_\epsilon} \quad t_\epsilon(x) = \sum_i w_{\epsilon,i}(x_i^2 - 1)$$

Notice again that under H_0 t_ϵ is a weighted sum of centered χ^2 random variables. When $\sum_i w_{\epsilon,i}^2 = 1/2$ we can appeal to the central limit theorem to say that t_ϵ is asymptotically $\mathcal{N}(0,1)$ when $\mathbb{E}[x] = 0$. The performance of the test is characterized by:

$$h_\epsilon = \inf_{v \in V_\epsilon} h_\epsilon(v, w_\epsilon) = \inf_{v \in V_\epsilon} \sum_i w_{\epsilon,i} v_i^2$$

The tests asymptotically distinguishes the null and alternative when $h_\epsilon \to \infty$ (Corollary 3.3).

A special case of this is when we set $w_{\epsilon,i} = \frac{1}{\sqrt{2n}}$ for $i \leq n$ and $w_{\epsilon,i} = 0$ otherwise. Then the performance is determined by the smallest projection onto the first n coordinates:

$$u_\epsilon = \frac{1}{\sqrt{2n}} \inf_{v \in V_\epsilon} ||\mathscr{P}_n(v)||_2^2$$

The χ^p**-tests** are extensions to the χ^2 tests using alternative powers. Define:

$$\psi_{\epsilon,T_\epsilon}^p = \mathbf{1}_{t_\epsilon > T_\epsilon} \quad t_\epsilon(x) \sum_i w_{\epsilon,i}(|x_i|^p - c_p)$$

where $c_p = \mathbb{E}[|\eta|^p]$ (recall $\eta \sim \mathcal{N}(0,1)$). As before, the performance is characterized by:

$$u_\epsilon^p = \inf_{v \in V_\epsilon} \sum_i w_{\epsilon,i} |v_i|^p$$

If $u_\epsilon^p \to \infty$ then the tests are minimax consistent (Proposition 3.2).

The χ^+**-test** is a combination of χ^2 and linear statistics:

$$t_\epsilon = \sum_i w_{\epsilon,i}(x_i^2 + x - 1)$$

Here we want non-negative weights that square-sum to 1/3 to have asymptotic normality when $\mathbb{E}[x] = 0$. Since $t + t^2 \geq t^p$ for $t \geq 0, p \in [1,2]$ this test is minimax consistent when (Proposition 3.3)

$$u_\epsilon^p = \inf_{v \in V_\epsilon} \sum_i w_{\epsilon,i} v_i^p \to \infty p \in [1,2]$$

The last test is the supremum test. Given a sequence $\{T_{\epsilon,i}\}_{i\in I}$ the supremum statistic is:

$$\psi_{\epsilon,T_\epsilon} = \mathbf{1}_{\mathscr{X}_\epsilon} \quad \mathscr{X}_\epsilon = \{x : \sup_i |x_i|/T_{\epsilon,i} > 1\}$$

Is based on the weighted infinity norm. Since the Type I error is $\alpha(\psi_\epsilon) = 1 - \Pi_i(1 - 2\Phi(-T_{\epsilon,i}))$ the only way to have $\alpha(\psi_\epsilon) \to 0$ is to have $\inf_i T_{\epsilon,i} \to \infty$. The Type II error (and consequently the performance) is governed by:

$$H_\epsilon(T_\epsilon, V_\epsilon) = \inf_{v\in V_\epsilon} \sup_i (|v_i - T_{\epsilon,i})$$

The supremum test is minimax consistent when $H_\epsilon \to \infty$ (Proposition 3.4).

As we did in the χ^2 case, we can consider projecting the sequence onto the first n components and then performing a supremum test. Here we should set the weights $T_{\epsilon,i} = \sqrt{2\log n}$ for $i \le n$ and 0 otherwise. We use $\sqrt{2\log n}$ here to ensure that the test has Type I error α (asymptotically). As you might expect the performance is governed by:

$$R_n(V_\epsilon) = \inf_{v\in V_\epsilon} ||\mathscr{P}_n(v)||_\infty$$

All of these tests can be defined and analyzed on the functional Gaussian model as well. See the book for details.

3.2 An Example: Spherical Alternative

When the alternative is seperated in ℓ_2 norm, we can use χ^2 tests and a bayes test based on a spherical alternative to establish minimax rates. In the sequence model, consider testing:

$$H_0 : v = 0 \quad H_1 = v :\in V_\epsilon(\rho_\epsilon) = \{v \in V_\epsilon : ||v||_2 \ge \rho_\epsilon\} \tag{14}$$

Here distinguishability is determined by the **inner radii**. For fixed n, define r_n to be the largest radius of an n-dimensional ball $D^n(r)$ that is contained in the set V_ϵ (not the set $V_\epsilon(\rho_\epsilon)$). Then define:

$$n(\rho_\epsilon) = \max\{n : r_n > \rho_\epsilon\}$$

While the ball $D^{n(\rho_\epsilon)}(r_{n(\rho_\epsilon)})$ is contained in V_ϵ by the definition of $n(\rho_\epsilon)$ we have that the *sphere* $S^{n(\rho_\epsilon)-1}(r_{n(\rho_\epsilon)})$ is completely contained in the lternative $V_\epsilon(\rho_\epsilon)$. Here is the intuition: the structure of V_ϵ will enforce that $r_n \to 0$ as $n \to \infty$. Think of V_ϵ as a Sobolev ball, so if we use more dimensions the ball will have to be smaller to accommodate the decaying coefficients. Then the parameter ρ_ϵ will identify a specific n for which the large sphere lies in the alternative.

We are now in an $n = n(\rho_\epsilon)$ dimensional problem and we can appeal to results from the finite dimensional setting. Specifically, we are now testing: $H_0 : v = 0$ from $H_1 = v \in V_\epsilon^n = \{v \in \mathbb{R}^n : ||v||_2 \ge \rho_\epsilon\}$ and the performance is governed by $u_\epsilon = \rho_\epsilon^2/\sqrt{2n}$.

To see why, let π be the uniform probability measure on the sphere $S^{n-1}(\rho)$ and by exploiting rotational invariance, we can rotate the observation X to $e_1||X||_2$ where e_1 is a standard basis element. The LRT is:

$$\begin{aligned} L_\pi(X) &= e^{-\rho^2/2}\int_{S^{n-1}(\rho)} e^{\langle v,X\rangle}\pi(dv) = C(n,\rho)\int_{-\rho}^{\rho} e^{t||X||_2}(\rho^2-t^2)^{(n-3)/2}dt \\ &= 2C(n,\rho)\int_0^{\rho}\cosh(||X||_2 t)(\rho^2-t^2)^{(n-3)/2}dt \end{aligned}$$

Which is an increasing function in $||X||_2$ so we we can just use the χ^2 test. In other words, $||X||_2$ is a sufficient statistic so it dictates both the upper and lower bounds. Under H_0, $||X||^2$ is a non-central χ^2 random variable (with CDF $G(x;t)$, t is the centrality parameter) with n degrees of freedom and we can appeal to Gaussian approximations of the distribution.

$$G_n(x,0) = \Phi((x-n)/\sqrt{2n}) + o(1) \quad G_n(x,t^2) = \Phi((x-n-t^2)/\sqrt{2n+4t^2}) + o(1)$$

which gives that $\rho^2/\sqrt{2n}$ governs the performance.

Example 4. *Consider the situation where V_ϵ is a Sobolev ellipsoid:*

$$V_{\epsilon,d} = \{v \in \ell^2 : \sum_i (v_i/d_i)^2 \le 1\}$$

and assume $d_i \to 0$, then $r_n = d_n$. In the special case of a power norm $d_i \asymp i^{-\eta}$ then $n(\rho_\epsilon) \asymp (\rho_\epsilon)^{-1/\eta}$ meaning that:

$$u_\epsilon \asymp (\rho_\epsilon)^{\frac{4\eta+1}{2\eta}}$$

At this point the actual rates depend on how the radius ρ_ϵ decays with ϵ.

To obtain upper bounds, we will use the notion of **Kolmogorov Diameters**. Define the accuracy of approximation of the set V_ϵ by the linear space L:

$$\delta(V_\epsilon, L) = \sup_{v\in V_\epsilon}\inf_{u\in L}||v-u||_2 = \sup_{v\in V_\epsilon}||v-\mathscr{P}_L(v)||_2$$

The Kolmogorov diameter R_n of dimension n is defined as:

$$R_n(V_\epsilon) = \inf_{L_n}\delta(V_\epsilon, L_n) \tag{15}$$

If we use an approximately best subspace, we will lose in our ability to approximate the signal, but in some cases we will be able to account for this loss. We will

Example 5. *Let V_ϵ be a Sobolev ellipsoid with d_i non-increasing. Then:*

$$R_n^2(V_\epsilon) = \sup_{v\in V_\epsilon}\sum_{i=n}^{\infty} v_i^2 \le d_n^2 \sum_{i=n}^{\infty}(v_i/d_i)^2 \le d_n^2$$

For the χ^2 test, projecting onto n dimensions the test is controlled by:

$$\frac{\rho_\epsilon^2 - R_n^2(V_\epsilon)}{\sqrt{2n}} = \frac{\rho_\epsilon^2 - d_n^2}{\sqrt{2n}}$$

And we can maximize over n. In the case of power norms $d_i \asymp i^{-\eta}$ this gives:

$$\sup_n \frac{\rho_\epsilon^2 - n^{-\eta}}{\sqrt{2n}} \asymp \rho_\epsilon^{\frac{4\eta+1}{2\eta}}$$

which establishes the minimax rate.

3.3 Constructing Priors

Recall that in testing the simple null hypothesis $v = 0$ against the bayesian hypothesis $v \sim \pi$, the performance of the test is governed by the likelihood ratio

$$L(x) = \frac{dP_\pi}{dP_0} = \int \exp(-||v||^2/2 + \langle x, v\rangle)\pi(dv)$$

and $\gamma(1)$ is determined by the L_1 or total variation distance. While this is hard to work with directly, we can appeal to Proposition **??** and work with ℓ_2 distance.

$$||P_\pi - P_0||_2^2 = \mathbb{E}_0[\mathbb{E}_\pi(dP_v/dP_0)]^2 - 1 = \int\int e^{\langle u,v\rangle}\pi(du)\pi(dv) - 1$$

We will frequently make use of **product priors** given by a sequence $\bar{\pi} = \{\pi_i, i \in I\}$ of probability measures on $\mathbb{R}^1$ defined as:

$$\pi(dv) = \Pi_i \pi_i(dv_i)$$

When the prior is a product prior, we can use write:

$$||P_{\bar{\pi}} - P_0||_2^2 \le \exp(\sum_i \int\int (e^{uv} - 1)\pi_i(du)\pi_i(dv)) - 1 \triangleq \exp(||\bar{\pi}||^2) - 1$$

We will call the term in the exponential the norm $||\bar{\pi}||^2$. We immediately have a proposition showing how $||\bar{\pi}||^2$ lower bounds the performance of the test.

Proposition 7 (Proposition 3.6)**.** *Suppose $\pi^\epsilon(V_\epsilon) \to 1$ where π^ϵ are the product priors corresponding to $\bar{\pi}_\epsilon$. If $||\bar{\pi}_\epsilon|| = o(1)$ then $\gamma(V_\epsilon) \to 1$ and if $||\bar{\pi}_\epsilon|| = O(1)$ then* $\liminf \gamma(V_\epsilon) > 0$.

It should be noted that the norm $||\bar{\pi}||$ on measures characterizes a hilbert space over measures over $\mathbb{R}$. That is the inner product $\langle \mu_1, \mu_2\rangle = \int\int e^{uv}\mu_1(du)\mu_2(dv)$ is positive definite and the associated norm is separating on the linear space of $\mathscr{L}$ of signed measures.

One useful prior is the product prior consisting of **two point factors**:

$$\pi(z,h) = (1-h)\delta_0 + h\delta_z \quad h \in [0,1], z \in \mathbb{R}^1$$

where δ_t is the dirac function. It is not too hard to show that:

$$||\bar{\pi}||^2 = \sum_i h_i^2(\exp(z_i^2) - 1)$$

And if $\sup_i |z_i| \leq B \to 0$ then by appealing to the relation $\exp(z^2) - 1 \approx z^2$ when $z \to 0$ we get:

$$\sum_i h_i^2 z_i^2 \leq ||\bar{\pi}||^2 \leq C(B) \sum_i h_i^2 z_i^2$$

for some constant $C(B) > 1, C(B) \to 1$ that depends only on B.

Another useful prior will be the product prior defined by symmetric **three-point factors**:

$$\pi(z,h) = (1-h)\delta_0 + h/2(\delta_{-z} + \delta_z) \;\; h \in [0,1], z \geq 0$$

Again it is not too hard to show that such a prior satisfies:

$$||\bar{\pi}||^2 = 2\sum_i h_i^2 \sinh^2(z_i^2/2)$$

And asymptotically (assuming $sup_i z_i \leq B \to 0, C(B) \to 1$ from above):

$$\sum_i h_i^2 z_i^4 \leq 2||\bar{\pi}||^2 \leq C(B) \sum_i h_i^2 z_i^4$$

Let us see where how these priors can be put to use via an example.

Example 6. *Consider testing $H_0 : v = 0$ from $H_1 : v \in V_n = \check{D}_{n,p}(\rho_n) = \{v \in \mathbb{R}^n : ||v||_p \geq \rho_n\}$ is the exterior of an ℓ_p ball. We would like to establish the minimax rate of:*

$$\rho^\star_{n,p} = \begin{cases} n^{(4-p)/4p} if & 0 < p \leq 2 \\ n^{1/2p} if & 2 < p < \infty \end{cases}$$

When $p \geq 2$ consider the product prior $\pi^n = (\pi(z_n, h_n))^n$ defined by two-point factors with $z_n = z > 1$ and $h_n = n^{-1}\rho_n^p = o(1)$. We first must verify that π^n concentrates on V_n. This can be done by applying the Chebyshev inequality to the random function $F_n = \sum_i v_i^p$. Then:

$$||\bar{\pi}_n||^2 \asymp n h_n^2 = \frac{\rho_n^{2p}}{n} = \left(\frac{\rho_n}{\rho_n^\star}\right)^{2p}$$

When $p \in (0,2]$, consider symmetric three-point factors with $z_n = n^{-1/p}\rho_n = o(1)$ and $h_n = 1$. Here it is obvious that π^n concentrates on V_n. Moreover

$$||\bar{\pi}_n||^2 = n z_n^4/2 = \left(\rho_n/\rho_n^\star\right)^4/2$$

The upper bounds follow from χ^p tests. If $p \leq 2$ we use the χ^2 test noting that:

$$n^{-1/2}||v||_2 \geq n^{-1/p}||v||_p \geq n^{-1/p}\rho_n \gg n^{-1/p}\rho_n^\star \geq n^{-1/4}$$

from which we note (recall u_n governed the performance of the χ^2 tests):

$$u_n = \frac{1}{\sqrt{2n}} \inf_{v \in V_n} \sum_i v_i^2 \to \infty$$

We can do the same thing with χ^p tests. This establishes $\rho_n^\star$ as the minimax rate.

To establish minimax rates for ℓ_∞ balls, we will need to use another type or prior. First we need some definitions. A collection of sequences $\bar{v}_\epsilon = \{v_{\epsilon,j}, j \in J_\epsilon\}$ with $v_{\epsilon,j} \in \ell^2$ is **orthogonal**, **semiorthogonal** or **asymptotically semiorthogonal** if for $k \neq j$ $\langle v_{\epsilon,j}, v_{\epsilon,k} \rangle = 0$, $\langle v_{\epsilon,j}, v_{\epsilon,k} \rangle \leq 0$ or $\sup_{k \neq j} \langle v_{\epsilon,j}, v_{\epsilon,k} \rangle = o(1)$. For a set of probability vectors $\bar{p}_\epsilon$ define the prior:

$$\pi^\epsilon(\bar{v}_\epsilon, \bar{p}_\epsilon) = \sum_j p_{\epsilon,j} \delta_{v_{\epsilon,j}}$$

These are called orthogonal, semiorthogonal, etc. depending on the properties of the collection $\bar{v}_\epsilon$. For asymptotically semiorthogonal priors we have:

$$||P_{\pi^\epsilon} - P_0||_2^2 \leq \sum_j p_{\epsilon,j}^2 (e^{|v_{\epsilon,j}||^2} - 1) + o(1)$$

Consequently, if the right hand side is $o(1)$ and the prior concentrates on the alternative then $\gamma(V_\epsilon) \to 1$ (Proposition 3.10).

We conclude this section with the example for $p = \infty$.

Example 7. *Consider testing $H_0 : v = 0$ from $H_1 = v \in V_n = \check{D}_{n,\infty}(\rho_n)$. We will show that: (1) if $\limsup \rho_n / \sqrt{\log n} < 1$ then $\gamma_n(V_n) \to 1$ and (2) if $\liminf \rho_n / \sqrt{\log n} > \sqrt{2}$ then $\gamma_n(V_n) \to 0$. For the lower bound, consider a uniform prior over the points $v_{n,j} = \rho_n e_j$ where e_j is the canonical basis in $\mathbb{R}^n$. We have:*

$$\sum_j p_{n,j}^2 \exp(||v_{n,j}||^2) = \frac{1}{n} \exp(\rho_n^2) = o(1)$$

For the upper bound, we apply the supremum test. There we say that we need to set the threshold to $\sqrt{2 \log n}$ to control the Type I error. With our setting for ρ_n, $\Phi(\sqrt{2 \log n} - \rho_n) \to 0$ establishing the minimax rate.

3.4 Classical Asymptotics and Triviality

We say a testing problem with null and alternative families $\mathscr{P}_0, \mathscr{P}_1$ is **trivial** if $\gamma(1, \mathscr{P}_0, \mathscr{P}_1) = 1$ (or $\to 1$). We say a problem has classical asymptotics if $\gamma(V(\rho)) \to 0$ as $\rho \to \infty$ where ρ parameterizes the separation.

A problem is trivial if and only if there exists a prior π_ϵ such that $||P_{\pi_\epsilon} - P_0||_1 \to 0$. In finite dimensional problems, triviality is possible if and only if the closure of the alternative V contains the point 0. In the infinite dimensional problem this is not the case as we have already seen.

Proposition 8 (Proposition 3.13)**.** *In infinite dimensional sequence model, the hypothesis testing problem $H_0 : v = 0$ against $H_1 : v \in V$ where $V = S_p(\rho)$ or $V = \check{D}_p(\rho)$ is trivial for any $\rho > 0, 0 < p \leq \infty$.*

One can prove this with the orthogonal prior that puts mass $1/n$ on each of the first n coordinates (scaled by ρ) and taking the limit as $n \to \infty$.

The situation is different if the alternative is a sobolev or power ellipsoid. Here we just state the power-norm result. Recall that $||v||_{\vec{a},p} = (\sum_i (v_i a_i)^p)^{1/p}$ and for the power norm we set $a_i = i^r$.

Theorem 9 (Theorem 3.1)**.** *In the testing problem with $H_1 : v \in V(\rho)$ where $V(\rho) = \check{D}_{r,p}(\rho) = \{v \in \ell^2 : ||v||_{r,p} > \rho\}$, set*

$$r_p^\star = \begin{cases} 1/4 - 1/p & \text{if} \quad p \le 2 \\ -1/2p & \text{if} \quad 2 < p < \infty \\ 0, & \text{if} \quad p = \infty \end{cases} \tag{16}$$

then the problem is trivial for $r \ge r_p^\star$ and has classical asymptotics when $r < r_p^\star$

The proof of lower bounds uses the product priors we have seen before: two point factors for $p \le 2$, three point factors for $2 < p < \infty$ and orthogonal collections for $p = \infty$. The upper bounds use the tests we have seen as well: χ^2 tests for $p \le 2$, χ^p tests for $2 < p < \infty$, and supremum tests for $p = \infty$.

One can show a similar triviality result for alternatives separated in Besov norm and alternatives corresponding to the intersection of an ℓ_p ellipsoid and the positive orthant. However in the last case the thresholds are slightly different, since, as we saw, for $0 < p \le 1$ the problem has convex alternative and we can use those results.

3.5 Nonclassical Asymptotics

In all of the results in the previos section, if $r \ge r_p^\star$ then we have no hope but if $r < r_p^\star$ the problem is trivial. In order to obtain nontrivial problems we will look at situations where $r \ge r_p^\star$ but place some additional constraints. This setting is called nonclassical.

One example we will focus on (in this document – not in the book) is the two-sided power norm constraint:

$$V(\kappa, \rho, R) = \{v \in \ell^2 : ||v||_{r,p} \ge \rho, ||v||_{s,q} \le R\}; \quad \kappa = (r, s, p, q) \tag{17}$$

The first question one must ask is when is the set V non-empty (See Corollary 3.15). Next, we can ask what setings of the parameters κ lead to trivial problems (See Theorem 3.5). To prove the lower bounds we carefully construct product priors defined earlier. The proof has many different cases since the settings of the parameters leading to triviality is somewhat complex and different techniques need to be used for different regions of that space.

The remainder of the section is devoted to proving analogous results for positive alternatives, two-sided besov norm alternatives (Section 3.4.7), and extended to the functional gaussian model (Section 3.4.8). As the techniques are similar, we refrain from diving into the details.

3.6 Rates for the Functional Gaussian Model

Recall the functional gaussian model:

$$dX_\epsilon(t) = s(t)dt + \epsilon dW(t), \quad s \in L_2(0,1)$$

And consider the alternative:

$$S_\epsilon = \{s \in L_2(0,1) : ||s||_p \geq r_\epsilon, ||s||_{(\eta,q)} \geq R\}$$

Where $||\cdot||_{(\eta,q)}$ is a Sobolev norm. The minimax rates in this problem are:

$$r_\epsilon^\star = \begin{cases} \epsilon^{4\eta/(4\eta+1)} & \text{if} \quad 1 \leq p \leq 2, q \geq p \\ \epsilon^{2\eta/(2\eta+1-1/p)} & \text{if} \quad 2 < p = q < \infty \\ (\epsilon^2 \log(1/\epsilon))^{\eta/(2\eta+1)} & \text{if} \quad p = q = \infty \end{cases} \tag{18}$$

Theorem 10 (Theorem 3.9)**.** *For $1 \leq p < \infty$ $\gamma(S_\epsilon) \to 1$ if $r_\epsilon = o(r_\epsilon^\star)$ and if $r_\epsilon / r_\epsilon^\star \to \infty$ then there are tests ψ_ϵ such that $\gamma(\psi_\epsilon, S_\epsilon) \to 0$. For $p = \infty$ there exists constants $\lambda_1 \leq \lambda_2$ such that $\gamma(S_\epsilon) \to 0$ as $\limsup r_\epsilon / r_\epsilon^\star < \lambda_1$ and $\gamma(S_\epsilon) \to 0$ as $\liminf r_\epsilon / r_\epsilon^\star > \lambda_2$.*

It is worth comparing the rates $r_\epsilon^\star$ to those required to estimate the signal $s(t)$. For finite p, $r_\epsilon^\star$ is asymptotically smaller than the estimation rate, meaning that we can detet these signals even when we are unable to estimate them. This is not too surprising. When $p = \infty$ the the detection and estimation rate are the same.

To prove the theorem we will reduce to the sequence gaussian model and use the tools we developed there. The lower bound proofs look like the usual nonparametric minimax proofs where we consider a prior based on small "bumps." The upper bounds use essentially histogram estimators.

For the lower bounds we need to fix a sufficiently smooth function $\phi(t)$ support on $(0,1)$ with $||\phi||_2 = 1$ and $||\phi^{(l)}||_q = C_{l,h} < \infty$ for $l \in [m+2]$. We will build priors which are based on the the functions $\phi_{n,i}(t) = \sqrt{n}\phi(nt - i)$ which have disjoint supports.

For $1 \leq p \leq 2$ consider the product priors based on symmetric three-point factors with $h = 1$, so that $\pi(z) = 1/2\delta_z + 1/2\delta_{-z}$. These priors are supported on the functions:

$$s_{\epsilon,\xi}(t) = \epsilon z_\epsilon \sum_i \xi_i \phi_{n,i}(t), \quad \xi_i \in \{\pm 1\}$$

This is the "hypercube" construction that we use in proving nonparametric lower bounds. At this point it is enough to check that π_ϵ concentrates on S_ϵ (the alternative) and then compute the norm $||\bar{\pi}_\epsilon||^2$. You can verify that π_ϵ is contained in the alternative when $n \asymp r_\epsilon^{-1/\eta}$ and $z_\epsilon \asymp \epsilon^{-1} n^{-(\eta+1/2)}$. The norm $||\bar{\pi}_\epsilon||^2 \asymp n_\epsilon z_\epsilon^4 \asymp (r_\epsilon / r_\epsilon^\star)^{4+1/\eta}$ which proves the lower bound. For $2 < p < \infty$ we again consider product priors based on symmetric three point factors but this time the prior is supported on:

$$s_{\epsilon,\xi} = \epsilon z_\epsilon \sum_i \xi_i \phi_{n,i}, \quad \xi = 0 \text{w.p.} 1 - n^{-1/2}, \pm 1 \text{w.p.} n^{-1/2}$$

Again one has to check the same conditions π_ϵ concentrates on S_ϵ and compute $||\bar{\pi}_\epsilon||^2$.

For $p = \infty$ we use uniform orthogonal priors concentrated on $s_{\epsilon,i} = \epsilon z_\epsilon \phi_{n,i}$.

The upper bounds are given by the usual nonparametric projection estimators. Fix $n = n_\epsilon \asymp (r_\epsilon^\star)^{-1/\eta}$ and consider an orthonormal collection $\bar{\phi}_{\epsilon,n}$ which are constant on the interval:

$$\phi_{\epsilon,i}(t) = \sqrt{n}\mathbf{1}[t \in [(i-1)/n, i/n]]$$

Then set:

$$X_{\epsilon,i} = \epsilon^{-1}\sqrt{n}\int_{(i-1)/n}^{i/n} dX_\epsilon(t)$$

We can now use χ^2, χ^p and supremum tests on the sequence $X_{\epsilon,i}$ with the appropriate choice of n to achieve the upper bounds.

4 Chapter 4

In this chapter, we establish a duality between asymptotically least favorable priors and asymptotically best tests. Finding both objects amounts to solving an extremal problem over a convex set in some Hilbert space.

4.1 Convex Alternative

In the gaussian sequence model, suppose the alternative $V \subset \ell^2$ is a convex set. Then we can show that the minimax performance is governed by:

$$u = \inf_{v \in V} ||v||$$

We can lower bound the performance as:

$$\beta(\alpha, V) \geq \sup_{v \in V} \beta(\alpha, \{v\}) = \sup_{v \in V} \Phi(\alpha - ||v||) \geq \Phi(\alpha - u)$$

In other words the asymptotically least favorable prior is a dirac mass at $v^\star$, which achieves the infimum above. The test statistic that achieves the minimax performance is the linear statistic $\psi_{r,T} = \mathbf{1}_{\langle x,r\rangle > T}$ for appropriately chosen threshold T. The performance of the linear statistic is:

$$\beta(\psi_{r,T}, V) = \sup_{v \in V} \Phi(T - \langle r, v\rangle)$$

Which is governed by:

$$h(r, V) = \inf_{v \in V} \langle r, v\rangle$$

The best direction r is obtained by solving a maximin problem:

$$r^\star = \text{argsup}_{r \in \ell^2 ||r|| \leq 1} \inf_{v \in V} \langle r, v\rangle$$

which is a dual problem to the extremal problem for u above. When V is convex, it is not too hard to show that $r^\star = v^\star/||v^\star||$, meaning that the performance of the linear statistic is governed by u.

This is the first example of the duality between lower bounds (the dirac prior) and upper bounds (the direction) in these problems.

Proposition 11 (Proposition 4.2)**.** *In testing $H_0 : v = 0$ against $H_1 : v \in V$ where V is a convex set of ℓ^2 with a minimum point $v^\star$ such that:*

$$u = ||v^\star|| = \inf_{v \in V} ||v|| > 0$$

Then:

$$\gamma(V) = 2\Phi(-u/2), \beta(\alpha, V) = \Phi(T_\alpha - u)$$

The proposition can be extended to the case where there is a convex set Z and a coordinate-wise one-to-one convex mapping $\phi : Z \to V$ given by $v_i = \phi_i(z_i) \geq 0$. This extension will be useful in "convexifying" sets.

Equipped with this result, we can establish sharp asymptotics for two sided constraints restricted to the positive orthant, where $p \leq 1$. The alternative is of the form:

$$V = \{v \in \ell^2_+ : ||v||_{r,p} \geq \rho, ||v||_{s,q} \leq R\}$$

where the norms are the power norms we saw earlier. When $q \geq 1$ the set is already convex. If $p \leq q < 1$ then we can convexify the set by the transformation $v_i = z_i^{1/p}$ and use the extension to Proposition 4.2. Obtaining the minimax performance now amounts to solving an extremal problem:

$$u^2 = \inf \left\{ \sum_i v_i^2 : v_i \geq 0, \sum_i i^{rp} v_i^p \geq \rho^p, \sum_i i^{sq} v_i^q \leq R^q \right\}$$

One can solve this problem via the method of lagrange multipliers for $p < q < \infty, p = q, p < q = \infty$ to arrive at the following theorem:

Theorem 12 (Theorem 4.1)**.** *Let $p \leq 1, p \leq q \leq \infty$ and consider:*

$$V_\epsilon = \{v \in \ell^2_+ : ||v||_{r,p} \geq \rho_\epsilon, ||v||_{s,q} \leq R_\epsilon\}, \;\; r \geq r^\star_{p,+} \triangleq 1/2 - 1/p$$

If $s - r \leq 1/p - 1/q$ then $\gamma(V_\epsilon) = 1$. If $s - r > 1/p - 1/q$ then as $R_\epsilon/\rho_\epsilon \to \infty$ one has:

$$\beta(\alpha, V_\epsilon) = \Phi(T_\alpha - u_\epsilon), \;\; \gamma(V_\epsilon) = 2\Phi(-u_\epsilon/2)$$

Where u_ϵ can be precisely computed for all cases.

Refer to the theorem in the book for details.

4.2 Product Priors and χ^2-tests

In the case where $p \le 2, q \ge 2$ we saw before that symmetric two-point prior are asymptotically least favorable and χ^2 tests are asymptotically minimax. We will establish a duality between finding the least favorable symmetric two-point prior and the best χ^2 test. This duality will allow us to pass to a similar extremal problem as we did in the previous section.

Recall that for a product prior $\pi(dv) = \Pi_i \pi_i(dv_i)$ we defined the norm $||\pi||$ to be:

$$||\pi||^2 = \sum_i ||\pi_i||^2 = \sum_i \int_{R^1} \int_{R^1} (e^{uv} - 1)\pi_i(du)\pi_i(dv)$$

We would like to show that:

$$L_\epsilon = \frac{dP_{\pi_\epsilon}}{dP_0} = \exp(-||\pi_\epsilon||^2/2 + ||\pi_\epsilon||\tau_\epsilon + \delta_\epsilon) \tag{19}$$

Where t_ϵ is asymptotically standard normal and $\delta_\epsilon \to 0$. The reason for showing this is that when $||\pi_\epsilon|| = O(1)$ we have:

$$\beta(\alpha; P_0, P_{\pi_\epsilon}) = \Phi(T_\alpha - ||\pi_\epsilon||) + o(1), \quad \gamma(P_0, P_{\pi_\epsilon}) = 2\Phi(-||\pi_\epsilon||/2) + o(1)$$

which follows (with some work) from the Neyman Pearson lemma.

Here we will establish Equation **??** for two-point and symmetric three-point factors. Recall the definitions:

$$\begin{aligned} \pi_{\epsilon,i} &= (1 - h_{\epsilon,i})\delta_0 + h_{\epsilon,i}\delta_{z_{\epsilon,i}} \\ \pi_{\epsilon,i} &= (1 - h_{\epsilon,i})\delta_0 + h_{\epsilon,i}(\delta_{z_{\epsilon,i}} + \delta_{-z_{\epsilon,i}})/2 \end{aligned}$$

Put:

$$u_{\epsilon,i} = ||\pi_{\epsilon,i}||, \quad u_\epsilon^2 = ||\pi_\epsilon||^2 = \sum_i u_{\epsilon,i}^2$$

The key property that we will use to establish Equation **??** is that particular subsequences of the product priors decay nicely. For any $\delta > 0$, denote:

$$J_{\epsilon,\delta} = \{i \in I : e^{2z_{\epsilon,i}^2} u_{\epsilon,i} > \delta\}, \quad u_{\epsilon,\delta}^2 = \sum_{i, J_{\epsilon,\delta}} u_{\epsilon,i}^2$$

And for now assume that $u_{\epsilon,\delta}^2 = o(1)$ for any δ. We will call this assumption A.2.

Proposition 13 (Proposition 4.4)**.** *If either $u_\epsilon = o(1)$ or $u_\epsilon \asymp 1$ and $u_{\epsilon,\delta}^2 = o(1)$ for any δ, then Equation **??** holds.*

Sketch. The case where $u_\epsilon = o(1)$ follows from the fact that $L_\epsilon \to 1$ in this case. In the other case, for any δ, we can decompose the likelihood ratio into terms in $J_{\epsilon,\delta}$ and the other terms. Using the fact that $u_{\epsilon,\delta}^2 \to 0$, we can essentially ignore the first terms. For the second set of terms, we use the fact that $e^{2z_{\epsilon,i}^2} u_{\epsilon,i} \le \delta$. Then taking a second order taylor expansion of the log-likelihood ratio, results in the three terms in Equation **??** □

It's best to see how this works with an example. Consider symmetric two-point priors π_ϵ with:

$$\pi_{\epsilon,i} = \pi(z_{\epsilon,i}, 1) = (\delta_{z_{\epsilon,i}} + \delta_{-z_{\epsilon,i}})/2$$

Assuming $\sup_i z_{\epsilon,i} = o(1)$ if $\sum_i z_{\epsilon,i}^4 = O(1)$ then assumption A.2 holds and we have:

$$||\pi_\epsilon||^2 \sim \sum_i z_{\epsilon,i}^4/2 \triangleq u_\epsilon^2(z_\epsilon)$$

So to find the best symmetric two-point product prior, we have to minimize $u_\epsilon^2(z_\epsilon)$ under the constraint $\pi_\epsilon(V_\epsilon) \to 1$. The constraint that the prior places its mass on the alternative can be re-written as:

$$u_\epsilon^2 = \frac{1}{2} \sum_{z_\epsilon \in V_\epsilon} \sum_i z_{\epsilon_i}^4 \tag{20}$$

And one can verify that our assumption holds as long as there exists a $z_\epsilon \in V_\epsilon$ such that $\frac{1}{2}\sum_i z_{\epsilon,i}^4 \sim u_\epsilon^2 + o(1)$. So this is our extremal problem for the lower bound.

As for the upper bound, recall the weighted χ^2-test $\psi_{\epsilon,w_\epsilon,T_\epsilon}(x) = \mathbf{1}[\sum_i w_{\epsilon,i}(x_i^2 - 1) > t_\epsilon]$ where the weights:

$$w_\epsilon \in W = \{w : w_i \geq 0, \sum_i w_i^2 = 1/2\}$$

The performance of the weighted χ^2-test is governed by:

$$h_\epsilon(w_\epsilon) = \inf_{v \in V_\epsilon} \sum_i w_{\epsilon,i} v_i^2$$

And if $h_\epsilon(w_\epsilon) = O(1)$, then $\beta(\alpha, V_\epsilon) \leq \Phi(T_\alpha - h_\epsilon(w_\epsilon)) + o(1)$ which we saw in Chapter 3. So to compute the best weights w_ϵ we must solve the maximin problem:

$$h_\epsilon = \sup_{w \in W} \inf_{v \in V_\epsilon} \sum_i w_{\epsilon,i} v_i^2 \tag{21}$$

Replacing $z_i^2/\sqrt{2}$ in Equation **??** with u_i, $v_i^2/\sqrt{2}$ in Equation **??** with u_i and $r_{\epsilon,i} = \sqrt{2} w_{\epsilon,i}$ the two problems become:

$$h_\epsilon = \sup_{r \in \ell_+^2, ||r|| \leq 1} \inf_{u \in U_\epsilon} \sum_i r_i u_i, \quad u_\epsilon = \inf_{u_\epsilon \in U_\epsilon} \sum_i u_{\epsilon,i}^2$$

Here U_ϵ is the image of the set V_ϵ under the mapping $u_i = v_i^2/\sqrt{2}$. This is the same primal-dual problem we saw in the previous section. In particular, if U_ϵ is convex, then $u_\epsilon = h_\epsilon$ governs the minimax performance. This is the content of Proposition 4.5.

We have just "convexified" the alternative. When $1 < p \leq 2$ the set $V_\epsilon = \{v \in \ell^2 : ||v||_p \geq \rho_\epsilon\}$ is non-convex, but for $p \leq 2$ the corresponding $U_\epsilon = \{u \in \ell_+^2 : \sum_i u_i^{p/2} \geq \sqrt{2}\rho_\epsilon^p\}$ is is convex. The same idea can be done for power norms and besov norms for the case $1 < p \leq 2$ to get sharp asymptotics in these settings. The main results are Theorem 4.2 and Proposition 4.8. In the functional gaussian model the main theorem is Theorem 4.4, which is an extension of Theorem 3.9 that we sketched above.

4.3 Degenerate Asymptotics

The last example of duality in this section comes from orthogonal priors and supremum tests. This leads to what the authors call asymptotic degeneracy.

Recall the orthogonal priors are formed by point masses at a collection of orthogonal vectors $\{v_{\epsilon,j}\}, j \in J_\epsilon$:

$$\pi_\epsilon = \pi(v_\epsilon, p_\epsilon) \sum_{j \in J_\epsilon} p_{\epsilon,j} \delta_{v_{\epsilon,j}}$$

given probability vectors p_ϵ. In this case, likelihood ratio is often not Gaussian as it was in the previous section and in fact we will show that it is "degenerate," meaning:

$$L_\epsilon = C_\epsilon + o(1)$$

Where $C_\epsilon \in [0,1]$ is non-random. If this is the case, then $\beta(\alpha) \geq (1-\alpha)(C_\epsilon + o(1)), \gamma(V_\epsilon) \geq C_\epsilon(1+o(1))$, which is what we refer to as asymptotic degeneracy.

When is L_ϵ degenerate? Let $u_{\epsilon,j} = ||v_{\epsilon,j}||$.

Proposition 14 (Proposition 4.10)**.** *Denote $u_\epsilon = \inf_j u_{\epsilon,j}$. Let D_ϵ be a family such that:*

$$w_\epsilon = u_\epsilon - D_\epsilon \to \infty, \sum_i e^{-w_{\epsilon,j}^2/2} \asymp 1, w_{\epsilon,j} = u_{\epsilon,j} - D_\epsilon$$

And set the probabilities p_ϵ to be:

$$p_{\epsilon,j} \propto e^{-w_{\epsilon,j}^2/2}$$

Then $L_\epsilon = \Phi(-D_\epsilon) + o(1)$ in P_0 probability.

Proof Sketch. Looking at the likelihood ratio, we can ignore some of the terms, looking instead at a truncated statistic, without significantly changing the likelihood ratio. By truncating the statistic appropriately, you can show that it has mean $\Phi(-D_\epsilon)$ and variance $\to 0$ which complete the proof. See the text for details. □

As for upper bounds, suppose the set V_ϵ has:

$$\inf_{v \in V_\epsilon} \sup_i (|v_i| - w_{\epsilon,i}) \geq D_\epsilon$$

where v_i are the coordinates of v in ℓ^2, for some D_ϵ and with w_ϵ satisfying:

$$\inf_j w_{\epsilon,j} \to \infty, \ \sum_j e^{-w_{\epsilon,j}^2/2} \asymp 1$$

then we have $\beta(\alpha, V_\epsilon) \leq (1-\alpha)(\Phi(-D_\epsilon) + o(1))$. We saw something like this in our analysis of orthogonal priors in chapter 3. On the other hand, if we take our orthogonal collection to be the standard basis vectors $\{u_{\epsilon,i} e_i\}$ and $w_{\epsilon,i} = u_{\epsilon,i} - D_\epsilon$ satisfies the above restrictions, then we have $\beta(\alpha, V_\epsilon) \geq (1-\alpha)(\Phi(-D_\epsilon) - o(1))$.

So for the upper bound we get to pick the thresholds $w_{\epsilon,j}$ and the performance is governed by:

$$\sup_{w} \inf_{v \in V_\epsilon} \sup_{i \in I} (|v_i| - w_{\epsilon,i})$$

while for the lower bound it is essentially governed by the vectors $u_{\epsilon,j}$:

$$\inf_{u : \sum_j u_{\epsilon,j} e_j \in V_\epsilon} \inf_j |u_{\epsilon,j}|$$

which again looks like a primal-dual pair.

Using this idea, you can characterize regions of the parameter space where two-sided alternatives (both power and besov norms) have asymptotic degeneracy.