The Philippine Statistician
1992, Vol. 41,
Pages 46-53

THE USE OF PROJECTION IN SEMIPARAMETRIC MODELS

Daniel Isidore Brian C. Bonzo
The Statistical Center
University of the Philippines Diliman

Abstract

Path analytic methods as developed by Johann Pfanzagl are used to demonstrate how estimator sequences can be improved. Improvement is shown through the projection of the estimator sequence. The projection technique is demonstrated in the case of parametric families and the corresponding extension to a special semiparametric family is given.

Keywords: Semiparametric models, tangent spaces, Hellinger distance, estimator sequence, projection

1. Introduction

Estimation is of basic interest in statistics. The general approach is as follows. A phenomenon is usually modeled by some probability measure P with distribution function F. It is characterized through functionals of the probability measure P, e.g., the mean and the variance. P is generally unknown and so are its functionals, hence the need for estimation.

In parametric theory, the probability measure is indexed by some parameter. In general, the form of the distribution is assumed to be known. The only missing information is the parameter value to completely describe the measure. Typical approaches to parametric estimation are maximum likelihood estimation (MLE), Least Square Estimation (LSE), Method of Moments Estimation (MME) and M-estimation. In nonparametric theory, the form of the probability measure is completely unknown. This makes it doubly hard to do estimation in this setting. Specialized techniques such as L-estimation and R-estimation, and the use of U-statistics are common in nonparametric procedures.

Recently, considerable interest in probability measures of mixed type has emerged. A probability measure of mixed type, is one wherein the form of the measure is partially known through the knowledge of an unknown indexing parameter but the distributional form is unknown. Oakes (1981) called such probability measures as semi-parametric models. An example is as follows. In a linear model setting, suppose the distributions of the error terms are identical but unknown. Then the density of Y_i, the dependent variable, is given by $n(y-\mathbf{x}_i'\mathbf{ß})$ where n is the unknown density of the errors, $\mathbf{x}_i$ is the set of regressors and **ß** is the regression coefficient. Here, the unknown density is indexed by the unknown coefficient **ß**.

Application of classical estimation procedures is difficult in the above setting. The usual approach is to construct asymptotically linear estimators which achieves the bound on the aymptotic variance. The key to the construction is the definition of the scores in the direction of the indexing parameter. If the parameter contains a nuisance parameter, the scores are projected onto the space of the parameters of interest to get the effective score.

In estimation, we usually have an apriori information about the value of the indexing parameter. This translates to estimation under a set of restrictions. In this context, we illustrate the use of projection in improving preliminary estimators, especially in the semiparametric setting.

Section 2 gives an introduction of tangent space and its role in estimation. Section 3 discusses distance measures between probability measures. Section 4 touches on projections of probability measures and section 5 contains the basic treatment for semiparametric models.

2. Tangent Spaces

Let $\mathcal{B}$ be a family of mutually absolutely continuous p-measure of some measurable space (Ω,F,P). The problem of estimation boils down to estimating the unknown measure $P\in\mathcal{B}$. Naturally, the candidate estimators are the P-measures Q which are "sufficiently" close to P. Identification of these measures requires knowledge of the local structure of $\mathcal{B}$ so as to assess the asymptotic performance of the measure Q.

The phrase "sufficiently close to" suggests distance and direction as criteria. It is convenient to regard the estimators Q as taking the form of a path denoted by, say, P_t. It is expected that for P_t to be reasonable, $P_t \to P$ as $t\to 0$. In asymptotics, P_t will be replaced by paths, like $P_{n\text{-}1/2}$ or $P_{n\text{-}1}$ so as to indicate that the estimator sequence is a function of n.

van der Vaart (1988) calls such paths P_t submodels of P. Furthermore, he defined a differential submodel P_t as a map $t\to P_t$ from [0,1] to $\mathcal{B}$ if there exists a measurable function g such that with $t\to 0$

$$\int\left[t^{-1}\left(p_t^{1/2} - p^{1/2}\right) - \tfrac{1}{2}gp^{1/2}\right]^2 du \to 0 \tag{2.1}$$

Here p_t and p are the μ-densities of P_t and P, respectively and μ is some measure dominating $\mathcal{B}$. Disregarding the integral in (2.1) yields g as a pointwise limit taking the form

$$\begin{aligned} g(x) &= 2p^{-1/2}\tfrac{d}{dt}p_t^{1/2}x\Big|_{t=0} \\ &= \tfrac{d}{dt}\log p_t x \end{aligned} \tag{2.2}$$

g can be considered, therefore, as a score function leading to P. Thus, the term directional derivative applies.

Pfanzagl (1982) gave a different definition of differentiable paths. A path $P_t\in\mathcal{B}$, $t\to 0$, is differentiable at $P\in\mathcal{B}$ with derivative g if the P-density of P_t can be expressed as

$$1+t(g+r_t)$$

where $g\in L_2(P)$ and $|r_t|\to 0$ as $t\to 0$. The two definitions are in fact, equivalent if we assume that g is dominated by a μ-square integrable function, say M. We present this in a form of a proposition.

Proposition: Let P_t be a path in $\mathcal{B}$ such that $P_t\to P\in\mathcal{B}$ as $t\to 0$. Let p_t and p be the μ-densities of P_t and P, respectively. Suppose there exists a function y which is dominated by some μ-square integrable function. Then P_t is differentiable if any of the following two conditions are satisfied.

(i) $p_t/p = 1 + t(g + r_t)$ where $\|r_t\|\to 0$ as $t\to 0$

(ii) $\int\left[t^{-1}\left(p_t^{1/2} - p^{1/2}\right) - \tfrac{1}{2}gp^{1/2}\right]^2 \to 0$ as $t\to 0$

Proof.: We show that (i) implies (ii) and then (ii) implies (i)

((i)⇒(ii)) Since $p_t = p(1 + t(g+r_t))$, we have by Taylor expansion and the Lebesgue dominated convergence theorem,

$$\int\left[t^{-1}\left(p_t^{1/2} - p^{1/2}\right) - \tfrac{1}{2}gp^{1/2}\right]^2 du = \int\left[t^{-1}\left(p^{1/2}\left(1 + \tfrac{1}{2}t(g + r_t) + o(t^2) - p^{1/2}\right) - \tfrac{1}{2}gp^{1/2}\right)\right]^2 du$$

$$= \int\left(\tfrac{1}{2}p^{1/2}r_t^{\circ} + o(t)p^{-1/2}\right)^2 du \to 0 \tag{2.3}$$

as $t \to 0$ since $|r_t| = o(t^0)$.

((ii)⇒(i)) If g is dominated by some μ-square integrable function we have, by the Lebesgue dominated convergence theorem,

$$\lim_{t\to 0}\left[t^{-1}\left(\frac{p_t^{1/2}}{p^{1/2}} - 1\right) - \tfrac{1}{2}g\right]^2 = 0 \tag{2.4}$$

This implies that for sufficiently small t we have, for every $\in > 0$

$$\left|\left(\frac{p_t^{1/2}}{p^{1/2}}\right) - \frac{1}{2}gt\right| < \in t \tag{2.5}$$

which implies

$$+t(g + r_t^-) \leq \frac{p_t}{p} \leq 1 + t(g + r_t^+) \tag{2.6}$$

where

$$r_t^- = -2t \in + t^2\left(\tfrac{1}{2}g - \in\right)^2$$
$$r_t^+ = 2t \in + t^2\left(\tfrac{1}{2}g + \in\right)^2$$

Note that r_t^- and $r_t^+ \to 0$ as $t \to 0$.

Hence the result.v

If $\mathcal{B}$ is a large set of probability measures we call the collection of all scores the tangent space of P and denote it by $T(P,\mathcal{B})$. The elements of $T(P,\mathcal{B})$ are used as approximations to the densities of p-measures in the neighborhood of P, with an approximation error becoming increasingly small as the measures near P.

Assumptions are made on $T(P,\mathcal{B})$ so that the estimator sequences for P will possess certain optimality properties. Assumptions like approximability, convexity, continuity and the requirement of $T(P,\mathcal{B})$ being a cone are common. For more details see Pfanzagl (1982).

As much as possible, we want the tangent space to be a linear space so that projections can be defined. In some cases, we require the tangent space to be full. Full in the sense that $T(P,\mathcal{B})$ defines the same space as

$$L^*(P) = \left\{g \in L_2(\mu):\ \int g d\mu = 0\right)$$

In such a case, estimator sequences obtained are necessarily asymptotically efficient. An example (see Pfanzagl, 1982) of a situation where we can get a full tangent space is the family of all p-measures Q equivalent to P with

$$\Delta(Q:P)^2 = \int \left({}^{q}\!/\!_{p} - 1 \right)^2 p d\mu \tag{2.7}$$

sufficiently small.

Example 1 In parametric families $\{P_\theta : \theta \in H), H \subset R^k)$ if $p^{(i)}(.,\theta)$: $(\cdot,\theta) := \frac{dp(\cdot,\theta)}{d\theta_i}$ fulfill for $i=1,\ldots,k$ some local Lipschitz condition then $T(P,\mathcal{B})$ = span $\{l^{(i)}(.,\theta), i=1,\ldots,k\}$ where $l^{(i)}(.,\theta) := \frac{p^{(i)}(.,\theta)}{p(\cdot,\theta)}$.

Example 2 In semiparametric models, the tangent space is given as follows. Let $B = \{P_{\theta,\tau} : \theta \in H , \tau \in T\}$ with $H \subset R$ (without loss of generality) and T some set (endowed with a topology). For a fixed θ, let $T_0(P_{\theta,\tau})$ denote the tangent space of $\{P_{\theta,\tau} : \tau \in T\}$ at $P_{\theta,\tau}$. Assuming regularity conditions on $\mathcal{B}$

$$T(P_\theta, \tau, P) = \left\{ c l^{(\cdot)}(\cdot, \theta, \tau) + h \colon c \in R,\ h \in \mathrm{T}_0 \ (P_\theta, \tau) \right\}$$

In short $T(P_{\theta,\tau}, \mathcal{B})$ is made up of direction derivatives orginating from two paths, namely, $P_{\theta_t,\tau} \to P_{\theta,\tau}$ and $P_{\theta,\tau_t} \to P_{\theta,\tau}$.

3. Distance Functions for Probability Measures

Having defined the set of possible directions towards the unknown measure P, it is important that we have a distance measure to assess how close our estimator sequence will be to the true measure P. Defining a distance measure for $\mathcal{B}$ provides a structure for our problem setting typical of metric spaçes. Hence, nice results from metric spaces can be borrowed to further characterize the estimation problem.

Let Q_1, Q_2, P belong to $\mathcal{B}$. A useful measure of distance between measures Q_1, Q_2 'near' P is given by

$$\Delta(Q_1, Q_2; P) = \left[\int \frac{(q_1 - q_2)^2}{p^2} \right]^{1/2} \tag{3.1}$$

If $Q_2 = P$ we get Pearson's distance measure

$$\Delta(Q;P)^2 = \int \left({}^{q}\!/\!_{p} - 1 \right)^2 p \, d\mu$$

This distance measure appears in a natural way in connection with likelihood ratios. A problem arises when one uses $\Delta(Q:P)$ as a distance measure for $\mathcal{B}$ since it is neither symmetric nor does it fullfill the triangle inequality. Thus, there is a need to find distance functions which are asymptotically equivalent to Δ.

An example of a distance measure which approximates Δ asymptotically is the Hellinger distance. The Hellinger distance is defined as

$$H^2(P,Q) \colon \ = 2 \int \left(q^{1/2} - p^{1/2} \right)^2 d\mu \tag{3.2}$$

It is related to the sup-distance

$$\mathrm{V}(Q,P) \colon \ = \sup(|Q(A) - P(A)| \colon A \in F) \tag{3.3}$$

through the inequality

$$\tfrac{1}{8} H(Q,P)^2 \leq \mathrm{V}(Q,P) \leq \tfrac{1}{2} H(Q,P)(1 - \tfrac{1}{16} H(Q,P))^{\frac{1}{2}} \tag{3.4}$$\

In connection with efficient estimation in semiparametric models, Begun, Hall, Huang and Wellner (1983) showed the importance of the root density $f^{1/2}(.,\theta,g)$ being Hellinger differentiable. They defined Hellinger-differentiability as follows. $f^{1/2}(.,\theta,g)$ is said to be Hellinger-differentiable at $(\theta,g) \in R \times G$ if there exists a function $r_\theta \in L_2(\mu)$ and a bounded linear operator $A : L_2(v) \to L_2(\mu)$ such that, with $f_n = (.;\theta_n,g_n)$,

$$\frac{\left\| f_n^{1/2} - f^{1/2} - \left\{ r_\theta(\theta_n - \theta) + A(g_n^{1/2} - g^{1/2} \right\} \right\|\mu}{|\theta_n - \theta| + \left\| g_n^{1/2} - g^{1/2} \right\|} \to 0 \text{ as } n \to \infty \tag{3.5}$$

for all sequence $\theta_n \to \theta$ and $g_n^{1/2} \to g^{1/2}$ in $L_2(v)$, where $g_n \in G$ for all $n \geq 1$, μ and v are Lebesgue measure in R^k; the expression in (3.5) is the differential at (θ,g).

<u>Example 3</u> For sequence of p-measures $P_{n\text{-}1/2}$ with P-density $1 + n^{-1/2} g + n^{-1/2} r_n$ with

$$E_p(r_n^2) = o(n^0)$$

then

$$H(P_{n-1/2}; P^n) = 8\left(1 - \exp\left[-\tfrac{1}{8} E_p(g^2)\right]\right) + o(n^o) \quad (P^n) \tag{3.6}$$

where P^n denotes a measure on the product space Ω^n and $E_p()$ the expectation with respect to P. Here H is a function of $g \in T(P,\mathcal{B})$. Thus, we see here explicitly the role of the tangent space in defining the desirability of using paths of the form $P_{n\text{-}1/2}$ in estimating the P-measure P.

4. Projections of Probability Measures

Estimation in our context will be dealt with in this way: some estimators for functionals of the probability measures will be obtained by finding first an estimator for the probability measure, say $P_n(\mathbf{x})$. Estimators for the functionals will then make use of $P_n(\mathbf{x})$ as the probability measure.

Our interest in projections is based on the following. Suppose we are given observations $X_1, \ldots X_n$ from X^n and we are given an estimator sequence $P_n(\mathbf{x}) \in \mathcal{B}$. If it is known that P belongs to some subfamily $\overline{\mathcal{B}} \subset \mathcal{B}$, it is then possible to obtain an estimator sequence $P_n'(\mathbf{x}) \in \overline{\mathcal{B}}$ based on $P_n(x)$. A way into this is through projection of $P_n(x)$ into $\overline{\mathcal{B}}$. Denote this projection by $P_n(\mathbf{x})$. In this way, we are able to obtain estimators which are strictly in. However, we expect that, under appropriate regularity conditions, the projection $P_n(\mathbf{x})$ improves $P_n(\mathbf{x})$.

We define a projection of a measure Q in this manner. $\overline{Q} \in \overline{\mathcal{B}}$ is a projection of Q into $\overline{\mathcal{B}}$ if q/q-1 is orthogonal to $T(\overline{Q},\overline{\mathcal{B}})$, i.e.,

$$E_{\overline{Q}}\left(\left(\tfrac{q}{\overline{q}} - 1\right)g\right) = 0 \quad \text{for all } g \in T(\overline{Q},\overline{B}) \tag{4.1}$$

or equivalently,

$$E_Q(g) = 0 \quad \text{for all g} \in T(\overline{Q},\overline{B})$$

Here q and $\overline{q}$ denote the densities of Q and $\overline{Q}$, respectively. It can be shown, under appropriate regularity conditions, that for any δ which approximates Δ, (see Pfanzagl, 1982)

$$\delta(Q,\overline{Q}) = \delta(Q,\overline{\beta}) + o(\delta(Q,\overline{Q})) \tag{4.2}$$

In short, $\overline{Q} \in \overline{\mathcal{B}}$ minimizes $\delta(Q,P)$ for any $P \in \overline{\mathcal{B}}$.

Let $1+k_Q$ and $1+k_{\overline{Q}}$ be the P-densities of Q and $\overline{Q}$, respectively. Then, under appropriate regularity conditions,

$$\begin{aligned} \left\| \bar{k}_Q - k_{\overline{Q}} \right\| &= \int \left(k_{\overline{Q}} - \bar{k}_Q \right)^2 p \, \mathrm{d}\mu \\ &= o(\Delta(Q;P)^2) \end{aligned} \tag{4.3}$$

where k_Q is the projection of $k_{\overline{Q}}$ into $T(P,\mathcal{B})$. This asserts that the projection of the density of Q is close to the density of the projected measure $\overline{Q}$ into $\mathcal{B}$, i.e., one may use $\bar{k}_Q$ instead of $\bar{k}_{\overline{Q}}$.

Since locally $\mathcal{B}$ behaves like a Hilbert space, we expect that the iterated projection results in Hilbert space theory apply to $\mathcal{B}$; i.e., let $\mathcal{B}_0 \subset \mathcal{B}_1$, and let Q be a p-measure not necessarily belonging to $\mathcal{B}_1$. Let Q_i denote the projection of Q into $\mathcal{B}_i$, i=0,1; then the projection $Q_{1,0}$ of Q_1 into $\mathcal{B}_0$ agrees closely with Q_0. Symbolically,

$$\Delta(Q_{1,0};Q_0) + o(\Delta(Q_{1,0};Q_0)) = \Delta(Q;Q_0)o(\Delta(Q_1;Q_0)^0) + \Delta(Q_1;Q_0)o(\Delta(Q_1;Q_0)^0) \tag{4.4}$$

Example 4 In parametric families $\mathcal{B} = \{P_\theta : \theta \in H\}$, $H \subset R^K$, the projection of Q (not in B) into $\overline{\mathcal{B}} \subset \mathcal{B}$, say $\overline{P}_\theta$, is determined by

$$E_Q(l^{(\cdot)}(\cdot,\underline{\theta})) = 0 \tag{4.5}$$

Under suitable regularity condition, $\underline{\theta}$ is determined by

$$\underline{\theta} - \theta = c(\theta,Q) + o(\Delta(Q;P_\theta)^2) \tag{4.6}$$

where

$$\begin{aligned} c(\theta,Q) &:= \left[\int l^{(\cdot)} l^{(\cdot)\prime} p \, \mathrm{d}\mu \right]^{-1} E_Q\left[l^{(\cdot)}(\cdot,\underline{\theta}) \right] \\ &= \Lambda(\theta) E_Q\left[l^{(\cdot)}(\cdot,\underline{\theta}) \right] \end{aligned} \tag{4.7}$$

It follows that $\overline{P}_\theta$ minimizes $\delta(Q,P_\tau)$ for $P_\tau \in \mathcal{B}$ up to a term of order $o(\delta(Q, P_\tau))$. Here δ is any distance function approximating Δ . It can be verified that the distance is minimized at τ given by

$$\tau = \theta + \Lambda(\theta) E_\theta\left[k_Q \, l^{(\cdot)}(\cdot,\theta) \right] \tag{4.8}$$

This is the improvement procedure form. Suitable estimators for τ can be arrived at by substituting "nice" estimators for θ, $\Lambda(\theta)$ and $E_\theta[K_Q \, l^{(\cdot)}(., \theta)]$.

5. A Projection Approach in Semiparametric Models Admitting a Sufficient Statistic

We now consider the case where the μ-density of P_t admits the representation

$$p(\cdot,\theta) = h(\cdot,\theta) g(\psi(\cdot,\theta)) \tag{5.1}$$

Here we say that $\Psi(.,\theta)$ is sufficient for the family $\{g \in T\}$ where T is endowed with some topology. From the previous section the one-step improvement form is given by

$$\tau = \theta + \Lambda(\theta) E_\theta \left[k_{P_n(x)} l^{(\cdot)}(\cdot,\theta) \right]$$

where $P_n(\mathbf{x})$ is the estimator sequence for P_θ. Under appropriate regularity conditions, the above sequence can be reduced to

$$\tau = \theta + \Lambda(\theta) E_{P_m} \left[l^{(\cdot)}(\cdot,\theta) \right] \tag{5.2}$$

We now construct an appropriate estimator for τ using P_n and some estimator $\hat{\theta}$ for θ.

The factorization of $p(.1,\theta)$ given above yields

$$\begin{aligned} l^{(\cdot)}(\cdot,\theta) &= \frac{h^{(\cdot)}(\cdot,\theta)}{h(\cdot,\theta)} + \frac{\psi(\cdot,\theta) g^{(\cdot)}(\cdot)}{g(\cdot,\theta)} \\ &= H(\cdot,\theta) + S(\cdot,\theta) \frac{g^{(\cdot)}(\cdot,\theta)}{g(\cdot,\theta)} \end{aligned} \tag{5.3}$$

where a dot on top indicates derivative with respect to θ. Given a sample of size n from $p(.,\theta)$, we split the sample and use the first K observations in constructing an estimator for θ.

Under appropriate regularity conditions (see Kumon and Amari, 1984), we can arrive at consistent estimators for θ using the estimating equations

$$\begin{aligned} \sum_{i=1}^{k} H(X_i,\theta) &= 0 \\ \sum_{i=1}^{k} S(X_i,\theta) &= 0 \end{aligned} \tag{5.4}$$

where $K < n$.

Using the rest of the sample, i.e. k+1,...,n we can use nonparametric techniques to construct consistent estimators for the functionals $\Lambda(\theta)$ and $E_{P_n}(p(.)(.,\theta))$. The construction of estimators is done by looking at the observations $\Psi_{K+1}(.,\theta), ..., \Psi_n(.,\theta)$ where θ, later on, takes the value of the solution to any one of the estimating equations given above. This procedure will enable us to get consistent estimator $g_n(.)$ for g((.,θ), the density of the sufficient statistic. Thus, we have an estimator for the improved estimator sequence $P\tau$.

References:

Begun, Hall, Huang and Wellner (1983). Information and asymptotic efficiency in parametric-nonparameteric Models. **Annals of Statistics 2,** 432-452.

Duncan, J. (1968). **The Elements of Complex Analysis**. Wiley, New York.

Kumon and Amari (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameter. **Biometrika 71,** 445-459.

Le Cam, L.(1986) **Asymptotic Models in Statistical Decision Theory**. Springer-Verlag, New York.

Oakes, D. (1981). Survival times: aspects of partial likelihood. **International Statistics Review 49,** 235-264.

Pfanzagl J. with Welfemeyer (1982) **Contributions to a General Asymptotic Statistical Theory**. Springer-Verlag, New York.

Pfanzagl J. (1990) **Estimation in Semiparametric Models**. Springer - Verlag, New York.

Rudin, W. (1964) **Principles of Mathematical Analysis**. McGraw Hill, New York.

Serfling, R. (1980) **Approximation Theorems in Mathematical Statistics**. Wiley, New York.

van de Vaart A.W. (1988). **Statistical Estimation in Large Parameter Spaces**. Thesis CWI tract 44, Centrum voor Wiskunde en Informatica, Amsterdam.