Semiparametric statistics

August 30, 2025

1 Heuristics

It is helpful to think about the theory at the heuristic level. Pfanzagl and von Mises posited the following expansion to hold for the functional of interest $\psi(P)$ where the first order Taylor expansion should be understood as an output of a linear operator

$$\psi(P_t) - \psi(P) = \dot{\psi}(P_t - P) + o(\|P_t - P\|) = \int \chi \mathrm{d}(P_t - P) + o(\|P_t - P\|),$$

taking limit $t \downarrow 0$, we have the arguably "Schrödinger equation" equivalent in statistics:

$$\frac{\mathrm{d}}{\mathrm{d}t = 0} \psi(P_t) = \int \chi \cdot s(x) \cdot p(x) \mathrm{d}x = \mathbb{E}(\chi \cdot s(X)),$$

subject to the constraint $\mathbb{E}\chi \equiv 0$. If the equation does not hold, it means that ψ is not sufficient smooth as a functional. A natural solution is to find an approximation of ψ, say $\widetilde{\psi}$ such that the above equation holds, and then control the approximation error. This route has been taken by **??**.

Why this equation is important? Suppose we replace P_t by $\widehat{P}_n$, some estimator of P:

$$\begin{aligned} \psi(\widehat{P}_n) - \psi(P) &= \int \chi \mathrm{d}(\widehat{P}_n - P) + o(\|\widehat{P}_n - P\|) \\ &= \int \chi \mathrm{d}\widehat{P}_n + o(\|\widehat{P}_n - P\|) \end{aligned}$$

suggesting that we can de-bias the "plug-in" estimator $\psi(\widehat{P}_n)$ by adding the influence function mean under $\widehat{P}_n$:

$$\widehat{\psi}_1 = \psi(\widehat{P}_n) - \int \chi \mathrm{d}\widehat{P}_n.$$

This heuristic also motivates the following, if you are concerned about the remainder term:

$$\psi(P_t) - \psi(P) = \int \chi_1(x) \mathrm{d}(P_t - P)(x) + \int\int \chi_2(x_1, x_2) \mathrm{d}(P_t - P)(x_1) \mathrm{d}(P_t - P)(x_2) + o(\|P_t - P\|^2)$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \psi(P_t) = \mathbb{E}(\chi_1(X) s_1(X)) + \mathbb{E}(\chi_2(X_1, X_2) s_2(X_1, X_2))$$

where

$$s_2(x_1, x_2) = \frac{1}{2} \left(\left. \frac{\mathrm{d}^2 p_t(x_1)}{\mathrm{d}t^2} \right|_{t=0} \frac{1}{p(x_1)} + \left. \frac{\mathrm{d}^2 p_t(x_2)}{\mathrm{d}t^2} \right|_{t=0} \frac{1}{p(x_2)} \right) + s_1(x_1) \cdot s_1(x_2).$$

If you are interested in this direction, please read the treatise **?**.

Recall the CRLB:

$$\sup_t \frac{(\frac{\mathrm{d}}{\mathrm{d}t=0}\psi(P_t))^2}{\mathbb{E}(s(X)^2)} = \sup_t \frac{\mathbb{E}^2(\chi \cdot s(X))}{\mathbb{E}(s(X)^2)} = \sup_{s \in \mathrm{T}_r(\mathcal{R})} \frac{\mathbb{E}^2(\chi \cdot s(X))}{\mathbb{E}(s(X)^2)} \leq \mathbb{E}(\chi^2)$$

$$\equiv \sup_{s \in \mathrm{T}_r(\mathcal{R})} I_s =: \text{semiparametric variance bound.}$$

There is a tension between the richness of the score (tangent spaces) and the richness of the influence functions: the richer the tangent space is, the less the influence functions there are.

2 Correspondence to differential geometry

In terms of the level that statisticians are using semiparametric theory, the value of more advanced differential geometry is somewhat limited. But if you are familiar with differential geometry, below is a mapping for you to gain deeper understanding. The kind of differential geometry that semiparametric theory is using is the so-called information geometry, initiated by Rao, Cox, Reid, Barndorff-Nielsen, Efron [**?**], and later by Amari, Jun Zhang, Battey (Cox's last student/post-doc) [**????**]. But possibly due to the pure-math nature, this field did not enter the mainstream statistical literature, nor the mainstream pure-math geometry literature, hence their own journal "Information Geometry". In particular, I personally really like Heather Battey's recent papers on this topic; she's doing some very specific cases but there could be a very general theme behind specific examples. But such works are more statistical in nature and cannot be easily replaced by AI.

Let $\mathcal{R} = \{r = \sqrt{p} : \|r\|^2 \equiv 1\}$ be the statistical manifold [**?**]. To each $r \in \mathcal{R}$, we attach a smooth curve $\gamma : \mathbb{R} \to \mathcal{R}$ such that $\gamma(0) = r$ and let $\gamma(t) \equiv r_t$ be the one-dimensional parametric submodel. To each $r \in \mathcal{R}$, define the velocity $\nu_{\gamma,r} : C^\infty(\mathcal{R}) \overset{\sim}{\to} \mathbb{R}$ as the homeomorphism

$$\nu_{\gamma,r} f = (f \circ \gamma)'(0)$$

for any chart $f : \mathcal{R} \to \mathbb{R}^d$. All such velocities constitute the so-called tangent space $\mathrm{T}_r\mathcal{R}$ locally at r. In statistics, the derivative operation is defined as the Derivative in Quadratic Mean (or Hellinger derivative) as follows:

$$\int \left(r_t - r - \frac{1}{2} t \cdot r \cdot s \right)^2 = o(t^2) \text{ or } \int \left(\frac{r_t - r}{t} - \frac{1}{2} r \cdot s \right)^2 = o(1), \text{ as } t \downarrow 0.$$

I want to highlight that the choice of DQM for differentiation is just one choice, not necessarily the only choice. For example, in differential privacy, there are results that use total variation distance as the metric [**??**].

The difficulty of using more differential geometry in semiparametric theory lies in the following unfortunate/fortunate fact: for nonparametric models, the statistical manifold is generally infinite-dimensional. Infinite-dimensional manifold is not an object extensively studied by pure mathematicians, except for a few special cases such as infinite-dimensional Lie group, and Wasserstein spaces (e.g. via the famous Otto calculus). One possible reason for this is the study of manifolds was motivated by physics – trying to describe the spacetime that we actually live in. But statistical manifolds equipped with Hellinger metric do not belong to that category.

3 Model tangent space

Most of the key references on semiparametric theory do not go into the rabbit hole on how to "compute" the tangent space corresponding to a statistical manifold $\mathcal{R}$. Generally, you can do the computation using the idea in the results below.

Theorem 1. *When we do not have any constraint on $\mathcal{R}$, $\mathrm{T}_r\mathcal{R} = L_0^2(r^2)$ has dimension ∞.*

Proof. For any function g such that $\mathbb{E}_r g \equiv 0$, $\mathbb{E}_r g^2 < \infty$ and $\mathbb{E}_r g^4 < \infty$ at $r \in \mathcal{R}$, we can establish the smooth curve $\gamma^2(t) : t \mapsto r^2(1 + t \cdot g) =: r_t^2$ such that: $\gamma(0) = r$ and

$$\begin{aligned}
\int \left(\frac{r_t - r}{t} - \frac{1}{2} r \cdot g \right)^2 &= \int \left(r \cdot \frac{\sqrt{1 + t \cdot g} - 1}{t} - \frac{1}{2} r \cdot g \right)^2 \\
&= \int \left(\frac{\frac{1}{2} t \cdot r \cdot g + \delta_t}{t} - \frac{1}{2} r \cdot g \right)^2 \\
&= \int \left(\frac{1}{2} (1 + \bar{t} \cdot g)^{-1/2} \cdot r \cdot g - \frac{1}{2} r \cdot g \right)^2 \\
&= \frac{1}{4} \int p g^2 \left\{ (1 + \bar{t} \cdot g)^{-1/2} - 1 \right\}^2 = o(1).
\end{aligned}$$

Hence $\overline{L_0^2(r^2) \cap L_0^4(r^2)} = \mathrm{T}_r(\mathcal{R})$. But $\overline{L_0^2(r^2) \cap L_0^4(r^2)} = L_0^2(r^2)$ so we prove the claim. □

Theorem 2. *For a log likelihood $\ell(x, y) = \ell(y|x) + \ell(x)$. The model tangent space can be decomposed into the following:*

$$\mathrm{T}_r(\mathcal{R}) = \Lambda_{Y|X} + \Lambda_X$$

and $\Lambda_{Y|X} \perp\!\!\!\perp \Lambda_X$.

Proof. The fact is trivial – any element in Λ_X is a measurable function of X, whereas any element in $\Lambda_{Y|X}$ is a measurable function of X, Y subject to the conditional mean given X equal to 0. □

Definition 3. The Efficient Influence Function is defined as

$$\mathsf{EIF} = \Pi[\mathsf{IF} | \mathrm{T}_r(\mathcal{R})]$$

where IF is any influence function and $\Pi[\cdot|\Lambda]$ is the L^2-projection operation onto the linear space Λ.

So the above definition tells us how to find the EIF – we first use our standard calculus of influence functions to find one influence function, then we find the tangent space and project.

Example 1. *Consider the following logistic regression:*

$$\mathbb{P}(A = 1|X) = \pi(X^\top \beta), \beta \in \mathbb{R}^d.$$

This model has conditional likelihood

$$
\begin{aligned}
&L_{A|X}(A,X;\beta) = \pi(X^\top\beta)^A(1-\pi(X^\top\beta))^{1-A} \Rightarrow \ell_{A|X}(A,X;\beta) = A\log\pi(X^\top\beta) + (1-A)\log(1-\pi(X^\top\beta)) \\
\Rightarrow s_{A|X}(A,X;\beta) &= \frac{A}{\pi(X^\top\beta)}\pi(X^\top\beta)(1-\pi(X^\top\beta))X - \frac{1-A}{1-\pi(X^\top\beta)}\pi(X^\top\beta)(1-\pi(X^\top\beta))X \\
&= \{A(1-\pi(X^\top\beta)) - (1-A)\pi(X^\top\beta)\}X = \{A-\pi(X^\top\beta)\}X.
\end{aligned}
$$

The part of the tangent space is then the closure of the span

$$\Lambda_{A|X} = \left\{\theta^\top\{A-\pi(X^\top\beta)\}X : \theta\in\mathbb{R}^d\right\}.$$

For X, since we do not impose anything on X, we have

$$\Lambda_X = \left\{h(X) : \mathbb{E}h(X)\equiv 0, \mathbb{E}h^2(X)<\infty\right\}.$$

Then $\mathrm{T}_r\mathcal{R} = \Lambda_{A|X} + \Lambda_X$.

Example 2. *Consider the distribution of X_1, X_2 without any constraint, which we call model (1):*

$$\ell(x_1,x_2) = \ell(x_1) + \ell(x_2|x_1).$$

Show the following: the tangent space of $p(x_1,x_2)$ is the direct sum between two L^2 spaces:

$$\mathrm{T}_p(\mathcal{P}) \equiv \Lambda_1 + \Lambda_{2|1}, \Lambda_1 = L_0^2(p_{X_1}), \Lambda_{2|1} = L_0^2(p_{X_2|X_1}) = \{g(X_1,X_2) : \mathbb{E}(g(X_1,X_2)|X_1)\equiv 0\}.$$

so $\mathrm{T}_p(\mathcal{P}) \equiv L_0^2(p_{X_1,X_2}) \equiv \left\{h(X_1,X_2) - \mathbb{E}h(X_1,X_2) : h(X_1,X_2)\in L^2(p_{X_1,X_2})\right\}$.

Now what if X_1 is independent of X_2, which we call model (2):

$$\ell(x_1,x_2) = \ell(x_1) + \ell(x_2)?$$

Then

$$\mathrm{T}_p(\mathcal{P}) \equiv \Lambda_1 \oplus \Lambda_2, \Lambda_1 = L_0^2(p_{X_1}), \Lambda_2 = L_0^2(p_{X_2}).$$

and now we have a nontrivial orthocomplement to $\mathrm{T}_p(\mathcal{P})$ because apparently $\mathrm{T}_p(\mathcal{P}) \neq L_0^2(p_{X_1,X_2})$ in this case. In particular we have the following:

$$
\begin{aligned}
\mathrm{T}_p(\mathcal{P})^\perp = &\ \left\{h(X_1,X_2)\in L_0^2(p(X_1,X_2)) : \mathbb{E}(h(X_1,X_2)|X_1) \equiv \mathbb{E}(h(X_1,X_2)|X_2)\equiv 0\right\} \\
= &\ \left\{h(X_1,X_2) - \mathbb{E}(h(X_1,X_2)|X_1) - \mathbb{E}(h(X_1,X_2)|X_2) + \mathbb{E}(h(X_1,X_2)) : h(X_1,X_2)\in L^2(p(X_1,X_2))\right\}.
\end{aligned}
$$

Derivation: take any $h_1\in L_0^2(p_{X_1})$ and any $h_2\in L_0^2(p_{X_2})$, for any $g\in\mathrm{T}_p(\mathcal{P})^\perp \subseteq L_0^2(\mu_{X_1,X_2})$ where we let μ denote the Lebesgue measure, we must have

$$
\begin{aligned}
&\mathbb{E}(g(X_1,X_2)h_1(X_1)) = \mathbb{E}(g(X_1,X_2)h_2(X_2)) = \mathbb{E}(g(X_1,X_2)) = 0 \\
\Rightarrow\ &\mathbb{E}(g(X_1,X_2)|X_1) = 0, \mathbb{E}(g(X_1,X_2)|X_2) = 0, \mathbb{E}(g(X_1,X_2)) = 0
\end{aligned}
$$

Whenever you obtain some influence function $\mathsf{IF}\equiv\mathsf{IF}(X_1,X_2)$, to obtain the EIF under model (2), we just project:

$$\mathsf{EIF} = \mathsf{IF} - \Pi[\mathsf{IF}|\mathrm{T}_p(\mathcal{P})^\perp] = \mathsf{IF} - (\mathsf{IF} - \mathbb{E}(\mathsf{IF}|X_1) - \mathbb{E}(\mathsf{IF}|X_2)) = \mathbb{E}(\mathsf{IF}|X_1) + \mathbb{E}(\mathsf{IF}|X_2).$$

Thought Experiment 4. In fact, the above results can be generalized to characterizing the model tangent space of any probability distribution Markov factorized according to a Directed Acyclic Graph (DAG) or Bayesian network (BayesNet) $\mathcal{G}(V, E)$ where V is the set of vertices and E is the set of directed edges without forming directed cycles:

$$p_{X_V}(x_V) = \prod_{v \in V} p_{X_v | X_{\mathsf{pa}_{\mathcal{G}}(v)}}(x_v | x_{\mathsf{pa}_{\mathcal{G}}(v)}),$$

where $\mathsf{pa}_{\mathcal{G}}(v)$ is the parent set of $v \in V$ relative to the DAG $\mathcal{G}$. For a complete DAG, i.e. every vertex is connected with every other vertex, the model tangent space is the entire $L_0^2(\mathbb{P}_{X_V})$. However, the Markov factorization property restricts that $X_v \perp\!\!\!\perp X_{\mathsf{nd}_{\mathcal{G}}(v)} | X_{\mathsf{pa}_{\mathcal{G}}(v)}$, $\mathsf{nd}_{\mathcal{G}}(v)$ denotes the non-descendants of $v \in V$ relative to the DAG $\mathcal{G}$. If there are missing edges, there will be algebraic (equality) constraints. Existence of equality constraints often points to a smaller tangent space. In certain models, there will be semi-algebraic (inequality) constraints. Existence of inequality constraints often does not reduce the tangent space [**?**]. If the model is finite-dimensional, "smaller" means smaller dimension. Such results belong to a field called "algebraic statistics", leading researchers including Robin Evans, Mathias Drton, Thomas Richardson and etc. from statistics, and Bernd Sturmfels, Cynthia Vinzant, June Huh (fields medalist) and etc. from pure math. It is at the intersection between probability theory and algebraic/tropical geometry.