Quick Notes on Manifold Fitting

Su Jiaji

October 25, 2024

This is a quick introduction of manifold fitting. More details can be found in:

- Fefferman, C., Ivanov, S., Kurylev, Y., Lassas, M., & Narayanan, H. (2018, July). Fitting a putative manifold to noisy data. *In Conference On Learning Theory* (pp. 688-720). PMLR.
- Yao, Z., & Xia, Y. (2019). Manifold fitting under unbounded noise. *arXiv preprint arXiv:1909.10228.*
- Yao, Z., Su, J., Li, B., & Yau, S. T. (2023). Manifold fitting. *arXiv preprint arXiv:2304.07680.*

Model Setting

Let $\mathcal{M}$ be a d-dimensional smooth latent manifold embedded in the ambient space $\mathbb{R}^D$. In this problem, we focus on a random vector $Y \in \mathbb{R}^D$ that can be expressed as

$$Y = X + \xi,$$

where $X \in \mathbb{R}^D$ is an unobserved random vector following a distribution ω supported on the latent manifold $\mathcal{M}$, and $\xi \sim \phi_\sigma$ represents the ambient-space observation noise, independent of X, with a standard deviation σ. The distribution of Y can be viewed as the convolution of ω and ϕ_σ, whose density at point y can be expressed as

$$\nu(y) = \int_{\mathcal{M}} \phi_\sigma(y - x)\omega(x)dx.$$

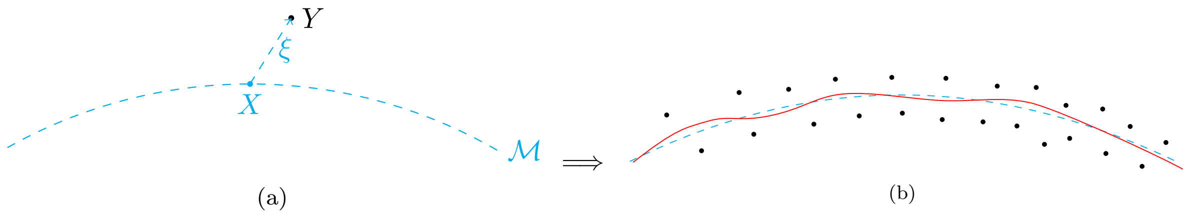

Figure 1: Illustration for the (a) model setting and (b) overall target, where the cyan parts are unknown/unobserved, black dots stand for the observations, and red curve represents the smooth d-dimensional manifold estimator we want.

Assume $\mathcal{Y}_N = \{y_i\}_{i=1}^N \subset \mathbb{R}^D$ is the collection of observed data points, also in the form of

$$y_i = x_i + \xi_i, \qquad \text{for } i = 1, \cdots, N,$$

with (y_i, x_i, ξ_i) being N independent and identical realizations of (Y, X, ξ). Based on $\mathcal{Y}_N$, we construct an estimator $\widehat{\mathcal{M}}$ for $\mathcal{M}$ and provide theoretical justification for it under the following main assumptions:

- The latent manifold $\mathcal{M}$ is a compact and twice-differentiable d-dimensional sub-manifold, embedded in the ambient space $\mathbb{R}^D$. Its volume with respect to the d-dimensional Hausdorff measure is upper bounded by V, and its *reach*[1] is lower bounded by a fixed constant τ.
- The distribution ω is a smooth distribution, with respect to the d-dimensional Hausdorff measure, on $\mathcal{M}$.
- The noise distribution ϕ_σ is a Gaussian distribution supported on $\mathbb{R}^D$ with density function

$$\phi_\sigma(\xi) = (\frac{1}{2\pi\sigma^2})^{\frac{D}{2}} \exp(-\frac{\|\xi\|_2^2}{2\sigma^2}).$$

- The intrinsic dimension d and noise standard deviation $\sigma < 1$ are known.

The manifold estimator $\widehat{\mathcal{M}}$ is suppose to be

- d-dimensional smooth manifold with lower bounded reach;
- close to $\mathcal{M}$.

[1]The value of reach($\mathcal{M}$) can be interpreted as a second-order differential quantity if $\mathcal{M}$ is treated as a function. Namely, for any arc-length parameterized geodesic γ of $\mathcal{M}$, $\|\gamma''(t)\|_2 \leq \text{reach}(\mathcal{M})^{-1}$ for all t.

Method

Let z be the point of interest, which is close to $\mathcal{M}$, and $z^* = \arg\min_{z' \in \mathcal{M}} d(z', z)$ be the projection of z on $\mathcal{M}$. Intuitively, the estimation of manifold can be viewed as "pushing" z to z^*. This pushing process involves two key components: direction and distance. The direction should be perpendicular to $T_{z*}\mathcal{M}$, which can be deduced from the local "covariance" structure, while the distance $d(z, \mathcal{M})$ might be estimated using the local average. The following subsections will introduce some intuitive concepts related to this process. For more details, please refer to the papers mentioned previously.

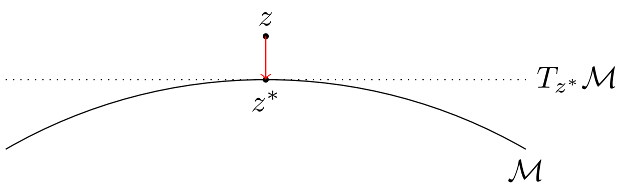

Figure 2: Push point z towards the manifold/z^*. The pushing should be perpendicular to $T_{z*}\mathcal{M}$.

Estimate Direction from Local "Covariance"

For each point $y_i \in \mathcal{Y}_N$, let y_i^* be its projection on $\mathcal{M}$. Assume the normal space of $\mathcal{M}$ at y_i^* has orthonormal basis $\{u_1, \ldots, u_{D-d}\}$, we use

$$\Pi_{y_i^*}^{\perp} = \big(u_1, \ldots, u_{D-d}\big)\big(u_1, \ldots, u_{D-d}\big)^{\top} = \sum_{k=1}^{D-d} u_k u_k^{\top}$$

to represent the projection matrix onto this space. This projection matrix can be estimated from the local variation centered at y_i, and the estimator of $\Pi_{y_i^*}^{\perp}$ is denoted as $\widehat{\Pi}_{y_i}^{\perp}$.

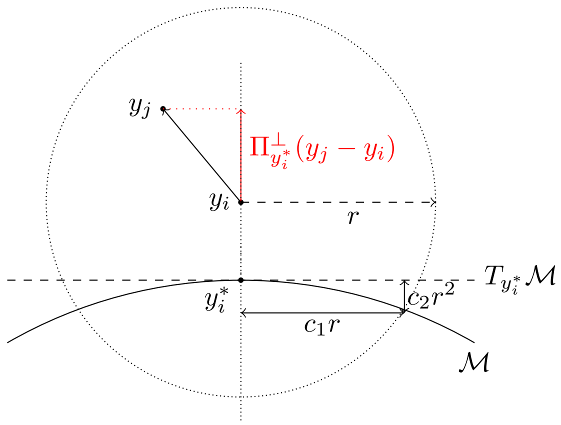

Figure 3: Illustration for the projection matrix and local variation.

Let $\mathcal{B}_D(y_i, r)$ be the D-dimensional Euclidean ball centered at y_i with radius r. If r is "large" enough, such that $\|y_i - y_i^*\| \le \|\xi_i\| \le c_0 r$, the area of $\mathcal{B}_D(y_i, r) \cap \mathcal{M}$ roughly has radius $c_1 r$, and the variation of $\mathcal{M}$ along the normal direction is less than $c_2 r^2$ due to the reach. Then, since the distribution of X is smooth, the variation of $Y - y_i$ along direction:

- $\leftrightarrow$: is roughly in the order of $(c_1 r)^2 + \sigma^2$;
- $\updownarrow$: is roughly bounded above by the order of $(c_2 r^2)^2 + \sigma^2$.

Thus, we can define

$$\widehat{\Sigma}_{r,i} = \frac{\sum_{j=1}^{N} (y_j - y_i)(y_j - y_i)^{\top} \mathbb{I}(\|y_j - y_i\| \le r)}{\sum_{j=1}^{n} \mathbb{I}(\|y_j - y_i\| \le r)}.$$

Then perform SVD on $\widehat{\Sigma}_{r,i}$ to obtain $\{\lambda_1 < \cdots < \lambda_D\}$ and $\{v_1, \ldots, v_D\}$, and estimate $\Pi^{\perp}_{y_i^*}$ with

$$\widehat{\Pi}^{\perp}_{y_i} = \big(v_1, \ldots, v_{D-d}\big)\big(v_1, \ldots, v_{D-d}\big)^{\top} = \sum_{k=1}^{D-d} v_k v_k^{\top},$$

whose estimation error can be bounded.

Smoothing System

To make the overall estimation smooth enough, the weight function for y_i with respect to z is defined as

$$\widetilde{\alpha}_i(z) = \left(1 - \frac{\|z - y_i\|^2}{r'^2}\right)^{\beta} \mathbb{I}(\|z - y_i\| \leq r'), \quad \alpha_i(z) = \frac{\widetilde{\alpha}_i(z)}{\sum_{i=1}^n \widetilde{\alpha}_i(z)},$$

where $\beta \geq 2$ is a parameter corresponding to the smoothness. Then, for z, a smooth reference point can be given by $\widehat{\mu}_z = \sum_{i=1}^N \alpha_i(z) y_i$, and a smooth projection matrix is calculated as

$$\Psi_z = \mathbb{P}_{D-d}\left(\sum_{i=1}^N \alpha_i(z) \widehat{\Pi}^{\perp}_{y_i}\right),$$

where $\mathbb{P}_k(A)$ stands for the projection of matrix A onto the span space corresponding to its largest k eigenvalues.

The Manifold Estimator

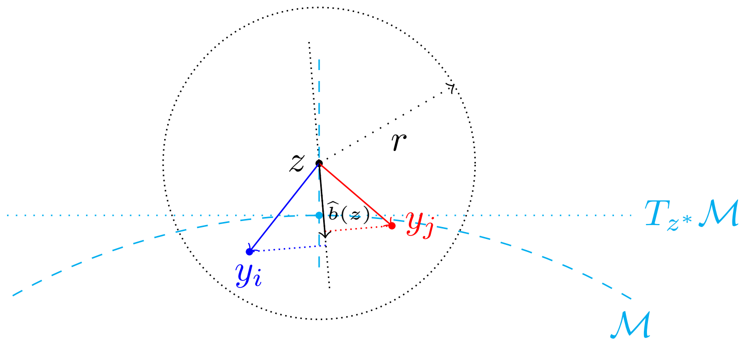

Figure 4: The bias vector.

The vector from z^* to z is estimated with the bias vector

$$\widehat{b}(z) = \sum_{i=1}^n \alpha_i(z) \Psi_z(z - y_i) = \Psi_z(z - \widehat{\mu}_z),$$

which can be shown

- $\|\widehat{b}(z)\|$ close to $\|z - z^*\|$;
- Jacobian matrix of $\widehat{b}(z)$ is close to Φ_z, i.e. $\|J_b(z) - \Psi_z\| \leq C\sigma/r' + o_p(1)$;
- Hessian matrix of $\widehat{b}(z)$ is lower bounded.

Finally, the manifold estimator is given by

$$\widehat{\mathcal{M}} = \{z \in \mathbb{R}^D : d(z, \mathcal{M}) < cr', \widehat{b}(z) = 0\}.$$

Under all the error bounds and all the smoothness, for any $z' \in \widehat{\mathcal{M}}$, with high probability,

- z' is close to $\mathcal{M}$;
- in its neighborhood, $\widehat{b}(z)$ is rank $D - d$.

Hence, with high probability, $\widehat{\mathcal{M}}$ is a d-dimension manifold, close to $\mathcal{M}$, and its reach can be bounded via the Hessian of $\widehat{b}(z)$.