Journal of Machine Learning Research 14 (2013) 1285-1302 Submitted 2/12; Revised 9/12; Published 5/13

Regularization-Free Principal Curve Estimation

Samuel Gerber SGERBER@MATH.DUKE.EDU
Mathematics Department
Duke University
Durham, NC 27708

Ross Whitaker WHITAKER@CS.UTAH.EDU
Scientific Computing and Imaging Institute
University of Utah
Salt Lake City, UT 84112

Editor: Sridhar Mahadevan

Abstract

Principal curves and manifolds provide a framework to formulate manifold learning within a statistical context. Principal curves define the notion of a curve passing through the middle of a distribution. While the intuition is clear, the formal definition leads to some technical and practical difficulties. In particular, principal curves are saddle points of the mean-squared projection distance, which poses severe challenges for estimation and model selection. This paper demonstrates that the difficulties in model selection associated with the saddle point property of principal curves are intrinsically tied to the minimization of the mean-squared projection distance. We introduce a new objective function, facilitated through a modification of the principal curve estimation approach, for which all critical points are principal curves and minima. Thus, the new formulation removes the fundamental issue for model selection in principal curve estimation. A gradient-descent-based estimator demonstrates the effectiveness of the new formulation for controlling model complexity on numerical experiments with synthetic and real data.

Keywords: principal curve, manifold estimation, unsupervised learning, model complexity, model selection

1. Introduction

Manifold learning is emerging as an important tool in many data analysis applications. In manifold learning, the assumption is that a given sample is drawn from a low-dimensional manifold embedded in a higher dimensional space, possibly corrupted by noise. However, the manifold is typically unknown and only observations from the ambient high-dimensional space are available. Manifold learning refers to the task of uncovering the low-dimensional manifold given the high-dimensional observations. Recent work in the machine learning community, such as Laplacian eigenmaps (Belkin and Niyogi, 2003), isomap (Tenenbaum et al., 2000), and locally linear embedding (Roweis and Saul, 2000) build directly on the manifold assumption and typically formulate the manifold estimation in terms of a graph embedding problem. The focus of these methods is to recover a low-dimensional description without modeling the manifold.

Hastie and Stuetzle (1989), motivated by the statistical properties of principal component analysis, introduced the notion of a principal curve passing through the *middle* of a distribution. The principal curve formulation casts manifold learning as a statistical fitting problem. Several approaches that fit a manifold model, such as autoencoder neural networks (Hinton and Salakhutdinov, 2006), self-organizing maps (Kohonen, 1988) and unsupervised kernel regression (Meinicke et al., 2005), can be implicitly or explicitly related to the principal curve formulation. The principal curve approach to manifold learning can be advantageous over alternatives that exclusively estimate descriptive parameters. For instance, the expected distance of the data to the manifold provides a quantitative measure of the geometric fit to the data. Additionally, the formulation provides a consistent approach to project and reconstruct unseen data points which can be important in practical applications.

While several methods have been proposed to estimate principal curves, a systematic approach to model selection remains an open question. To facilitate the discussion, recall the formal definition of principal curves. Let Y be a random variable with a smooth density p such that the support $\Omega = \{y : p(y) > 0\}$ is a compact, connected region with a smooth boundary.

Definition 1 (Principal Curve, Hastie and Stuetzle, 1989) *Let* $g : \Lambda \to \mathbf{R}^n$, $\Lambda \subset \mathbf{R}$ *and* $\lambda_g : \Omega \to \Lambda$ *with* projection index $\lambda_g(y) = \max_s\{s : \|y - g(s)\| = \inf_{\tilde{s}}\|y - g(\tilde{s})\|\}$. *The principal curves of* Y *are the set* $\mathcal{G}$ *of smooth functions* g *that fulfill the self consistency property* $E[Y|\lambda_g(Y) = s] = g(s)$.

Hastie and Stuetzle (1989) showed that principal curves are critical points of the expected projection distance $d(g,Y)^2 = E[\|g(\lambda_g(Y)) - Y\|^2]$, that is, for g a principal curve the (Fréchet) derivative of $d(g,Y)^2$ with respect to g is zero. Estimators for principal curves typically minimizes $d(g,Y)^2$ directly over a suitable representation of g (e.g., kernel smoother, piece-wise linear). However, Duchamp and Stuetzle (1996) showed, for principal curves in the plane, that all critical points are saddles. Thus, there are always adjacent (nonprincipal) curves with lower projection distance.

In fact, without additional regularization, principal curve estimation results in space-filling curves, which have both lower training and test projection distance than curves that provide a more *meaningful* summary of the distribution. Thus, all estimation approaches impose, implicit or explicit, regularity constraints on the curves (Tibshirani, 1992; Kégl et al., 2000; Chang and Ghosh, 2001; Smola et al., 2001; Meinicke et al., 2005) which, due to the saddle point property, will lead to estimates that lie on the constraint boundary and typically do not satisfy the conditions for principal curves. Using only the data, there is no well-defined, *correct* amount of regularization, and a regularization strategy would necessarily depend on extraneous considerations (e.g., the particular properties of the application). Thus, Duchamp and Stuetzle (1996) noted: “To our knowledge, nobody has yet suggested a reasonably motivated, automatic method for choice of model complexity in the context of manifold estimation or nonparametric orthogonal distance regression. This remains an important open problem”. In this paper we propose a solution to address the model complexity challenge in principal curve estimation.

This paper contends that the model selection problem is intrinsically tied to the mean-squared-error objective function of the principal curve formulation. In the supervised setting, for example, for regression, the mean squared error provides an adequate measure for model selection (Härdle and Marron, 1985). In the unsupervised setting, for manifold estimation, the missing information on variable relationships renders distance measures inadequate. To be more precise, in regression for two observations with the same predictor measurements, the differences in predicted value is necessarily due to noise. Manifold estimation faces an inherent ambiguity in whether differences between observations should be treated as variation *orthogonal* to the manifold or as variation *along* the manifold, as illustrated in Figure 1.

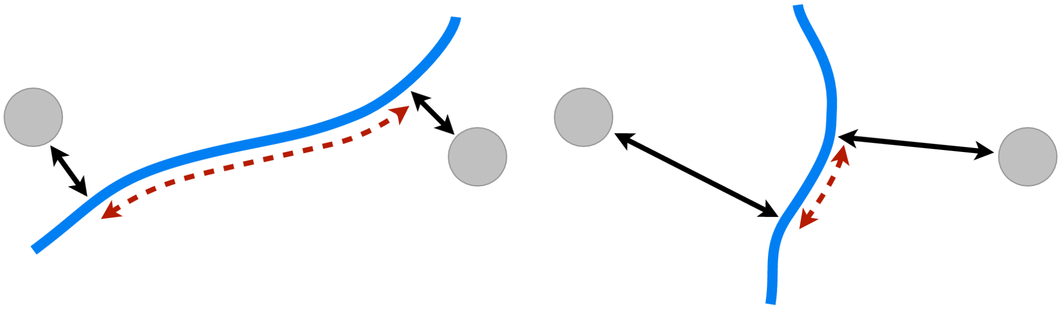

Figure 1: Two curves (solid lines) with a different trade off between variation along (dashed arrows) and orthogonal to (solid arrows) the curve for two data points in identical positions. A direct minimization of mean squared projection distance $d(g,Y)^2 = E[\|g(\lambda(Y)) - Y\|^2]$ will always favor curves passing close to the data points and not necessarily move towards a principal curve.

Informally, this indicates why principal curves are not minima of the distance function; trading off variation orthogonal to the manifold with variation along the curve leads to curves with smaller projection distances. This becomes evident in cross-validation, where a test example has no fixed assigned manifold coordinate (as compared to the fixed predictor variables in regression) but has manifold coordinates assigned based on the current fit. This leaves the flexibility to adapt the manifold coordinates of the test data to minimize projection distance, and leads, almost always, to simultaneous decreases in train and test error. This would not pose a significant problem if the principal curves would be minima of the mean squared projection distance—one would rely on the local minimum provided by the optimization. However, due to the saddle point property, the optimization moves towards a curve that passes through all the training data and cross-validation does not provide any reliable indication for early stopping. Thus, principal curve estimation based on minimizing the projection distance is not a viable path for automatic model selection.

Recent research produced some sophisticated approaches to automatically select the regularization parameters, such as the minimum description length approach by Wang and Lee (2006) and the oracle inequality approach by Biau and Fischer (2012). However, the fundamental problem persists; a regularized principal curve, is estimated which will always force the principal curve conditions to be violated. To expand on this point, consider this comparison to the regression setting. For nonparametric least squares regression, the regularization is a mechanism to cope with finite sample sizes. If the complete density is known, the regularization is unnecessary. The regression curve that minimizes the least squares objective function, the conditional expectation given the predictor, is the desired solution. Thus, with increasing sample size the amount of regularization required decreases to zero, for example, the bandwidth in kernel regression. For principal curves, the regularization is required even if the full density is available; the regularization serves as a crutch to fix a shortcoming in the objective function. This paper proposes an approach to principal curve estimation that fixes this shortcoming in the objective function. The proposed alternative objective function, for which

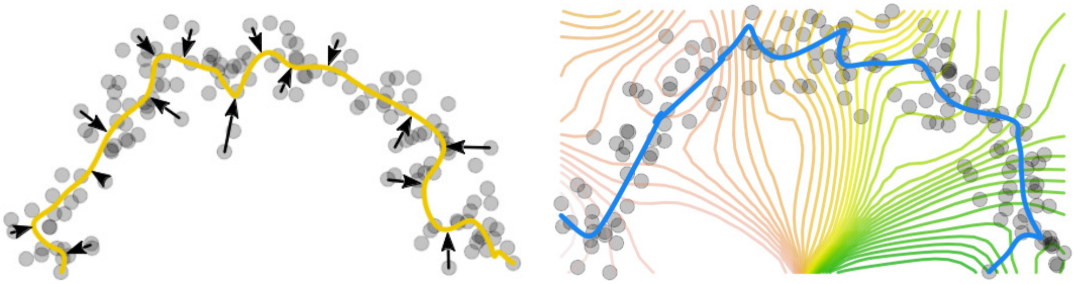

Figure 2: Illustration of the traditional approach compared to the proposed method for estimating principal curves. (left) The traditional approach specifies a curve g (yellow) and defines λ_g as the coordinate corresponding to the closest point on the curve. (right) The proposed formulation specifies and arbitrary λ—here a few level sets of λ are shown—which in turn defines the curve $g = E[Y|\lambda(Y)]$ (blue). In the traditional approach g is optimized to find a principal curve. In the new formulation the optimization is over λ.

principal curves emerge naturally as minima, alleviates the need for regularization or changes of the original principal curve definition.

The proposed formulation rests on the observation that minimizing $d(g,Y)^2$ allows principal curves g to move away from saddle points and improve their fit by violating the self consistency property. This leads to the idea of turning the optimization problem around. Instead of having the projection index λ_g controlled by the curve g, we fix $g \equiv E[Y|\lambda(Y)]$ to the conditional expectation given λ and let λ be an arbitrary function $\lambda : \Omega \mapsto \mathbf{R}$ (with mild conditions, to be specified in Section 2). To distinguish from Hastie and Stuetzle (1989), with λ_g fixed given g, we call λ the *coordinate mapping* and $E[Y|\lambda(Y)]$ the *conditional expectation curve* of λ. The difference between the two formulations is illustrated in Figure 2. Now, the strategy is to optimize over the coordinate mapping λ rather than g. Minimizing $d(\lambda,Y) = E\left[\|g(\lambda(Y)) - Y\|^2\right]$ with respect to λ leads to critical points that are, indeed, principal curves, but unfortunately they are still saddle points. In this case, escaping from saddle points is possible by violating the orthogonal projection condition. This inspires the design of an objective function, facilitated through this level-set based formulation, that penalizes nonorthogonality of the coordinate mapping λ

$$q(\lambda,Y)^2 = E\left[\left\langle (g(\lambda(Y)) - Y)\,,\, \frac{d}{ds}g(s)\Big|_{s=\lambda(Y)} \right\rangle^2\right] .$$

Here, g is by definition self-consistent and the orthogonal projection of λ is enforced by the objective function. In this formulation, all critical points λ^* are minima with $q(\lambda^*,Y)^2 = 0$. Hence, the conditional expectation curve at a minima $g \equiv E[Y|\lambda^*(Y)]$ corresponds to a principal curve (see Section 2).

The proposed formulation leads to a practical estimator, that does not impose any model complexity penalty in its estimate. This regularization-free estimation scheme has the remarkable property that a minimum on training data typically corresponds to a desirable solution (see Sections 3 and 4).

2. Conditional Expectation Curves

Let Y be a random variable with a smooth density p with support $\tilde{\Omega} = \{y : p(y) > 0\} \subset \mathbf{R}^n$ such that the closure $\Omega = \mathrm{cl}(\tilde{\Omega})$ is a compact, connected region with smooth boundary. This includes densities p that smoothly go to 0 at the boundary with an open support as well as densities with a compact support. Let $\lambda : \Omega \to \Lambda$ be Lipschitz with Λ a connected closed subset of $\mathbf{R}$. To shorten notation, let μ be the measure such that p is the Radon-Nikodym derivative of μ with respect to the Lebesgue measure and write $d\mu = d\mu(y) = \mu(dy) = p(y)dy$.

As for principal curve setting, the conditional expectation $E[Y|\lambda(Y)]$ is defined over a set of measure zero and is not a priori well defined. To formally define the conditional expectation let

$$g_\sigma(s) = \frac{\int_\Omega K_\sigma(s - \lambda(y))y d\mu}{\int_\Omega K_\sigma(s - \lambda(y))d\mu}$$

with $K_\sigma(s)$ a mollifier. Now we define the conditional expectation as the limit

$$\begin{aligned} E[Y|\lambda(Y) = s] = \lim_{\sigma \to 0} g_\sigma(s) &= \frac{\int_\Omega \delta(\lambda(y) - s)y d\mu}{\int_\Omega \delta(\lambda(y) - s)d\mu} \\ &= \frac{\int_{\lambda^{-1}(s)} y p(y) J_\lambda(y)^{-1} d\mathcal{H}_{n-1}(y)}{\int_{\lambda^{-1}(s)} p(y) J_\lambda(y)^{-1} d\mathcal{H}_{n-1}(y)} \, . \end{aligned} \qquad (1)$$

with $J_\lambda(y)$ the 1-dimensional Jacobian of λ and $\mathcal{H}_{n-1}$ dimensional Hausdorff measure. The second equality follows from the uniform convergence of the mollifier to the Dirac delta distribution. The last equality invokes the coarea formula (Federer, 1969) by changing to integration over the level sets of λ. This definition is, for λ an orthogonal projection, equivalent to the conditional expectation as understood in Hastie and Stuetzle (1989) and Duchamp and Stuetzle (1996). This formulation also directly extends to define conditional expectation manifolds with $\Lambda \subset \mathbf{R}^m$.

Depending on λ and Ω the conditional expectation (1) can be discontinuous, for example, for λ with constant regions or for λ and Ω that result in *abrupt* topological or geometrical changes in the level sets. However, additional conditions on λ and Ω, ensure that g is a differentiable.

Lemma 2 *For $\lambda \in C^2(\Omega)$, g is differentiable Λ-almost everywhere.*

Proof The derivative of g is

$$\frac{d}{ds} g(s) = \frac{\frac{d}{ds} m(s)}{n(s)} - g(s) \frac{\frac{d}{ds} n(s)}{n(s)} \, .$$

with

$$\begin{aligned} m(s) &= \int_{\lambda^{-1}(s)} y p(y) J_\lambda(y)^{-1} d\mathcal{H}_{n-1}(y) \, , \\ n(s) &= \int_{\lambda^{-1}(s)} p(y) J_\lambda(y)^{-1} d\mathcal{H}_{n-1}(y) \, . \end{aligned}$$

By Sard's theorem, the set of critical values of λ has measure zero in Λ, and by the implicit function theorem, $\lambda^{-1}(s)$ is a Riemannian manifold Λ-almost everywhere. Thus, integrating with respect to the Hausdorff measure is equivalent to integration over the manifold $\lambda^{-1}(s)$, Λ-almost everywhere. Then, by the general Leibniz rule (Flanders, 1973), the derivatives $\frac{d}{ds}m(s)$ and $\frac{d}{ds}n(s)$ are

$$\frac{d}{ds}\int_{\lambda^{-1}(s)}\omega = \int_{\lambda^{-1}(s)} i(\mathbf{v}, d_y\omega) + \int_{\partial\lambda^{-1}(s)} i(\mathbf{v},\omega) + \int_{\lambda^{-1}(s)}\frac{d}{ds}\omega \tag{2}$$

with $i(\cdot,\cdot)$ the interior product, d_y the exterior derivative with respect to y and ω the differential $(n-1)$-form $yp(y)J_\lambda(y)^{-1}d\lambda^{-1}(s)$ for $\frac{d}{ds}m(s)$ and $p(y)J_\lambda(y)^{-1}d\lambda^{-1}(s)$ for $\frac{d}{ds}n(s)$, respectively. The vector field $\mathbf{v}$ is the change of the manifold $\lambda^{-1}(s)$ with respect to s. For Ω with smooth boundary $\partial\Omega$, the restriction $\lambda|_{\partial\Omega}$ is also in $C^2(\partial\Omega)$. Thus, by the implicit function theorem, $\mathbf{v}$ exist Λ-almost everywhere for both λ and $\lambda|_{\partial\Omega}$ and it follows that $\frac{d}{ds}g(s)$ exists Λ-almost everywhere. ■

Corollary 3 *If $\lambda \in C^2(\Omega)$, $J_\lambda(y) > 0$ Ω-almost everywhere and $J_{\lambda|_{\partial\Omega}}(y) > 0$, $\partial\Omega$-almost everywhere then g is differentiable everywhere.*

The conditions in Corollary 3 ensure that the integrals in Equation (2) are well defined for all $s \in \Lambda$ and hence, the derivative of g is defined everywhere as well. As for the conditional expectation (1), Lemma 2 and Corollary 3 do not rest on a scalar λ and directly extends to conditional expectation manifolds.

With the definition of $g \equiv E[Y|\lambda(Y)]$, the conditional expectation given a coordinate mapping λ, one is tempted to minimize the distance function

$$d(\lambda, Y)^2 = E\left[\|g(\lambda(Y)) - Y\|^2\right]$$

with respect to λ. This formulation does, indeed, lead to conditional expectations at critical points that are principal curves and the self consistency cannot be violated by definition. Unfortunately, critical points are still saddles. In this case, λ can move towards conditional expectation curves with lower objective by violating the orthogonal projection requirement (rather than violating the self consistency requirement, as happens in the original formulation). However, the switch from optimization of curves g to optimization of coordinate mappings λ permits an alternative objective function that penalizes nonorthogonal λ:

$$q(\lambda, Y)^2 = E\left[\left\langle g(\lambda(Y)) - Y, \left.\frac{d}{ds}g(s)\right|_{s=\lambda(Y)}\right\rangle^2\right]. \tag{3}$$

For $q(\lambda, Y)$ to be well defined $\frac{d}{ds}g(s)$ has to exist Ω-almost everywhere and thus, requires the conditions of Corollary 3.

The next Section shows that all critical points of (3) are minima and that the corresponding conditional expectation curves are principal curves in the sense of Hastie and Stuetzle (1989).

2.1 Properties of Critical Points

The following definition introduces a slightly weaker version of principal curves. For principal curves which have no ambiguity points, that is, all $y \in \Omega$ have a unique closest point on the curve, the definition is equivalent to principal curves.

Definition 4 (Weak Principal Curves) *Let $g : \Lambda \to \mathbf{R}^n$ and $\lambda : \Omega \to \Lambda$. The weak principal curves of Y are the set $\mathcal{G}_w$ of functions g that fulfill the self consistency property $E[Y|\lambda(Y) = s] = g(s)$ with λ satisfying $\left\langle y - g(\lambda(y)),\, \frac{d}{ds}g(s)\big|_{s=\lambda(y)} \right\rangle = 0 \forall y \in \Omega$.*

Weak principal curves do include all principal curves but can potentially also include additional curves for λ_g with discontinuities, that is, curves with ambiguity points. However, under restriction to continuous λ, which excludes principal curves with ambiguity points, the set of principal curves and weak principal curves are identical under this restriction. Furthermore, in practical applications principal curves with ambiguity points are typically not of interest.

Theorem 5 *The conditional expectation curves of critical points of $q(\lambda, Y)^2$ are weak principal curves.*

Proof The self-consistency property follows immediately from the definition of g as conditional expectation given λ. Thus, to show that critical points of $q(\lambda, Y)^2$ are principal curves, requires $\frac{d}{ds}g(s)$ to be orthogonal to $g(s) - y$ almost everywhere for y in the level set $\lambda^{-1}(s)$.

Let τ be an admissible variation of λ, that is, $\tau \in C^3$ and satisfies the conditions of Corollary (3). Taking the Gâteaux derivative of $q(\lambda, Y)^2$ with respect to τ yields

$$\frac{d}{d\varepsilon} q(\lambda + \varepsilon\tau, Y)^2\bigg|_{\varepsilon=0} = \int_\Omega \left\langle (g(\lambda(y)) - y)\,,\, \frac{d}{ds}g(s)\bigg|_{s=\lambda(y)} \right\rangle \\ \frac{d}{d\varepsilon}\left\langle (g(\lambda(y) + \varepsilon\tau(y)) - y)\,,\, \frac{d}{ds}g(s)\bigg|_{s=\lambda(y+\varepsilon\tau(y))} \right\rangle\bigg|_{\varepsilon=0} d\mu\,.$$

It immediately follows λ is critical if it is orthogonal to its conditional expectation curve Ω-almost everywhere.

To exclude other possible critical points requires that there is no λ for which

$$\left\langle \frac{d}{d\varepsilon} g(\lambda(y) + \varepsilon\tau(y))\bigg|_{\varepsilon=0}\,,\, \frac{d}{ds}g(s)\bigg|_{s=\lambda(y)} \right\rangle + \\ \left\langle (g(\lambda(y)) - y)\,,\, \frac{d}{d\varepsilon}\frac{d}{ds}g(s)\bigg|_{s=\lambda(y+\varepsilon\tau(y))}\bigg|_{\varepsilon=0} \right\rangle = 0$$

for all admissible τ.

For the variation of $g(\lambda(y))$ take the Gâteaux derivative of the mollified expectation

$$\frac{d}{d\varepsilon} g_\sigma(\lambda(y) + \varepsilon\tau(y))\bigg|_{\varepsilon=0} = \frac{\int_\Omega K'_\sigma(\lambda(y) - \lambda(\tilde{y}))(\tau(y) - \tau(\tilde{y}))\big(\tilde{y} - g_\sigma(\lambda(y))\big)d\mu(\tilde{y})}{\int_\Omega K_\sigma(\lambda(y) - \lambda(\tilde{y}))d\mu(\tilde{y})}$$

with $K'_\sigma(s - \lambda(\tilde{y})) = \frac{d}{ds}K_\sigma(s - \lambda(\tilde{y}))$. Separating the terms into $\tau(y)$ and $\tau(\tilde{y})$ yields

$$\frac{d}{d\varepsilon} g_\sigma(\lambda(y) + \varepsilon\tau(y))\bigg|_{\varepsilon=0} = \tau(\lambda(y))\,\frac{d}{ds}g(s)\bigg|_{s=\lambda(y)} + \alpha_\sigma(\lambda(y), \tau)$$

with

$$\alpha_\sigma(s, \tau) = \frac{\int_\Omega K'_\sigma(s - \lambda(\tilde{y}))\tau(\tilde{y})\,(\tilde{y} - g_\sigma(s))\,d\mu}{\int_\Omega K_\sigma(s - \lambda(\tilde{y}))d\mu}$$

The same procedure for $\frac{d}{ds}g(s)$ yields

$$\begin{aligned}\frac{d}{d\varepsilon}\left.\left.\frac{dg_\sigma(s)}{ds}\right|_{s=\lambda(y)+\varepsilon\tau(y)}\right|_{\varepsilon=0} =& \frac{\int_\Omega K''_\sigma(\lambda(y)-\lambda(\tilde{y}))(\tau(y)-\tau(\tilde{y}))\,(y-g_\sigma(\lambda(y)))\,d\mu(\tilde{y})}{\int_\Omega K_\sigma(\lambda(y)-\lambda(\tilde{y}))d\mu(\tilde{y})}\\ &-\frac{\int_\Omega K'_\sigma(\lambda(y)-\lambda(\tilde{y}))(\tau(y)-\tau(\tilde{y}))d\mu}{\int_\Omega K_\sigma(\lambda(y)-\lambda(\tilde{y}))d\mu(\tilde{y})}\left.\frac{d}{ds}g_\sigma(s)\right|_{s=\lambda(y)}\\ &-\left.\frac{d}{ds}g_\sigma(s)\right|_{s=\lambda(y)}\left.\frac{d}{d\varepsilon}g(\lambda(y)+\varepsilon\tau(y))\right|_{\varepsilon=0}\end{aligned}$$

and by separating the terms

$$\frac{d}{d\varepsilon}\left.\left.\frac{dg_\sigma(s)}{ds}\right|_{s=\lambda(y)+\varepsilon\tau(y)}\right|_{\varepsilon=0} = \tau(y)\left.\frac{d^2}{ds^2}g_\sigma(s)\right|_{s=\lambda(y)} + \beta_\sigma(\lambda(y),\tau)$$

with

$$\begin{aligned}\beta_\sigma(s,\tau) = &-\frac{\int_\Omega K''_\sigma(s-\lambda(\tilde{y}))\tau(\tilde{y})\,(y-g_\sigma(s))\,d\mu}{\int_\Omega K_\sigma(s-\lambda(\tilde{y}))d\mu}\\ &+\frac{\int_\Omega K'_\sigma(s-\lambda(\tilde{y}))\tau(\tilde{y})d\mu}{\int_\Omega K_\sigma(s-\lambda(\tilde{y}))d\mu}\frac{d}{ds}g_\sigma(s) - \frac{d}{ds}g_\sigma(s)\alpha_\sigma(s,\tau)\,.\end{aligned}$$

Terms of the form $\int_\Omega K'_\sigma(s-\lambda(\tilde{y})f(\tilde{y})d\tilde{y}$ converge uniformly to $\int_\Omega \delta'(s-\lambda(\tilde{y})f(\tilde{y})d\tilde{y} = \frac{d}{ds}\int_{\lambda^{-1}(s)} f(\tilde{y})d\tilde{y}$ which under the conditions of Lemma (2) exist. Thus, $\frac{d}{ds}g_\sigma(s)$ converges uniformly and with the uniform convergence of $g_\sigma \to g$ yields the limit $\lim_{\sigma\to 0}\frac{d}{ds}g_\sigma(s) = \frac{d}{ds}g(s)$. With $\alpha(s,\tau) = \lim_{\sigma\to 0}\alpha_\sigma(s,\tau)$ and $\beta(s,\tau) = \lim_{\sigma\to 0}\beta_\sigma(s,\tau)$ and taking the limit as $\sigma \to 0$ yields

$$\begin{aligned}&\frac{d}{d\varepsilon}\left.\left\langle (g(\lambda(y)+\varepsilon\tau(y))-y)\,,\,\left.\frac{d}{ds}g(s)\right|_{s=\lambda(y+\varepsilon\tau(y))}\right\rangle\right|_{\varepsilon=0}\\ =&\tau(y)\left\|\left.\frac{d}{ds}g(s)\right|_{s=\lambda(y)}\right\|^2 + \left\langle \left.\frac{d}{ds}g(s)\right|_{s=\lambda(y)},\alpha(\lambda(y),\tau)\right\rangle\\ &+\left\langle g(\lambda(y))-y,\left(\tau(y)\left.\frac{d^2}{ds^2}g(s)\right|_{s=\lambda(y)}+\beta(\lambda(y),\tau)\right)\right\rangle.\end{aligned}$$

Note that

$$\int_{\lambda^{-1}(s)}(g(s)-y)\,d\mu = g(s)\int_{\lambda^{-1}(s)}d\mu - \int_{\lambda^{-1}(s)}y d\mu = 0\,,$$

and thus the term $\left\langle \alpha(\lambda(y),\tau),\, \frac{d}{ds}g(s)\big|_{s=\lambda(y)}\right\rangle$ with constant value along the level sets of λ does not contribute to the variation. Therefore, a critical point for which

$$\left\langle g(\lambda(y))-y,\,\left.\frac{d}{ds}g(s)\right|_{s=\lambda(y)}\right\rangle \neq 0$$

Ω-almost everywhere, there must be a λ for which

$$\left\|\left.\frac{d}{ds}g(s)\right|_{s=\lambda(y)}\right\|^2 + \left\langle g(\lambda(y))-y,\,\left.\frac{d^2}{ds^2}g(s)\right|_{s=\lambda(y)}+\beta(\lambda(y),\tau)\right\rangle = 0$$

Ω-almost everywhere. Since $\beta(\lambda(y),\tau)$, $\frac{d}{ds}g(s)\big|_{s=\lambda(y)}$ and $\frac{d^2}{ds^2}g(s)\Big|_{s=\lambda(y)}$ are constant along the level set of λ requires that $\left\| \frac{d}{ds}g(s)\big|_{s=\lambda(y)} \right\|^2 = 0$ and either $g(\lambda(y)) - y = 0$ or

$$\left\langle g(\lambda(y)) - y, \frac{d^2}{ds^2}g(s)\bigg|_{s=\lambda(y)} + \beta(\lambda(y),\tau) \right\rangle = 0\,,$$

Ω-almost everywhere. It follows that for all critical points

$$\left\langle g(\lambda(y)) - y, \frac{d}{ds}g(s)\bigg|_{s=\lambda(y)} \right\rangle = 0\,.$$

■

Corollary 6 *All critical points of $q(\lambda, Y)^2$ are minima.*

Proof By definition $q(\lambda,Y)^2 \geq 0$ and by Theorem 5 $q(\lambda,Y)^2 = 0$ for all critical points. ■

Corollary 7 *The coordinate mapping λ is a minima of $q(\lambda,Y)$ if and only if $E[Y|\lambda(Y)]$ is a weak principal curve.*

Proof If λ is minimal it follows from Theorem 5 that $E[Y|\lambda(Y)]$ is a weak principal curve. If $E[Y|\lambda(Y)]$ is a principal curve λ has to be an orthogonal projection and thus $q(\lambda,Y)^2 = 0$ is minimal. ■

2.2 Remarks

The minima include *pathological* conditional expectation sets such as single points, curves with discontinuities and space-filling curves. However, if Ω and p admit a minimal λ that satisfies the condition in Lemma 2, the conditional expectation is a well behaved curve Λ-almost everywhere.

In principle, $q(\lambda,Y)^2$ can be minimized by shrinking the tangent $\frac{d}{ds}g(s)$, that is, through a reparametrization of g. If λ is constrained so that the resulting conditional expectation has a fixed-length tangent, for example, a unit parametrization, or if the objective function is changed to

$$q_n(\lambda,Y)^2 = E\left[\frac{\left\langle g(\lambda(Y)) - Y, \frac{d}{ds}g(s)\big|_{s=\lambda(Y)} \right\rangle^2}{\left\| \frac{d}{ds}g(s)\big|_{s=\lambda(Y)} \right\|^2} \right]\,,$$

then shortening the tangent does not change the objective. However, in practice, using the approach described in Section 3, this problem did not arise and the empirical differences between minimizing q_n and q were negligible.

At critical points (local minima), the length of the tangent $\frac{d}{ds}g(s)$ has no effect on either the objective function or the solution. Changing the coordinate mapping to $\tilde{\lambda}(y) = \eta(\lambda(y))$ with η :

$\Lambda \mapsto \mathbf{R}$ differentiable and monotonic results in a re-parametrization of the conditional expectation curve, thus, the minimal sets consist of an infinite number of curves. Only for a re-parametrization $\eta(s) = c$, that is, a constant λ, is the geometry of the curve changed to a single point. Thus, the set of local minima of this objective form a single connected set, with many re-parametrizations and different curves connected at the solution $\lambda = c$. In terms of the objective function $q(\lambda, Y)$, these are all equally good solutions. Similar to principal components, choosing among the possible principal curves an alternate criterion, possibly application dependent, is required. An obvious choice is the mean squared projection distance and exclude the trivial space filling solution with zero mean squared projection distance.

3. An Algorithm for Data Sets

Gerber et al. (2009) proposed an algorithm to minimize $d(\lambda, Y)^2$ using a radial basis function expansion for λ and a kernel regression for the conditional expectation g. For a data set $Y = \{y_i\}_1^n$, the radial basis function expansion is

$$\lambda(y) = \frac{\sum_{j=1}^n K(y - y_j) z_j}{\sum_{j=1}^n K(y - y_j)}$$

with K a Gaussian kernel with bandwidth σ_λ and $Z = \{z_i\}_1^n \in \mathbf{R}$ a set of parameters. The parameters Z are a set of locations in Λ, the range of the coordinate mapping λ. Each point in Ω is thus mapped to a weighted average of the parameters Z. The expected squared projection distance $d(\lambda, Y)^2$ is estimated with the sample mean-squared projection distance $\hat{d}(\lambda, Y)^2 = \frac{1}{n} \sum_{i=1}^n \| g(\lambda(y_i)) - y_i \|^2$. Gerber et al. (2009) employ a gradient descent on Z to minimize $\hat{d}(\lambda, Y)^2$. Here, the same approach is used for minimizing

$$\hat{q}(\lambda, Y) = \frac{1}{n} \sum_{i=1}^n \left\langle g(\lambda(y_i)) - y_i, \frac{d}{ds} g(s) \Big|_{s=\lambda(y_i)} \right\rangle^2 \tag{4}$$

but instead of a Nadaraya-Watson kernel regression (Nadaraya, 1964; Watson, 1964) for g we use a locally linear regression (Cleveland, 1979), for better behavior for data near the boundaries of Λ. The algorithm is implemented in R in the package *cems* (Gerber, 2011).

The gradient of (4) is

$$\begin{aligned} \frac{d}{dz_k} \hat{q}(\lambda, Y)^2 = \sum_{i=1}^n & \left\langle (g(\lambda(y_i)) - y_i), \frac{d}{ds} g(s) \Big|_{s=\lambda(y_i)} \right\rangle \\ & \left[\left\langle \frac{d}{dz_k} g(\lambda(y_i)), \frac{d}{ds} g(s) \Big|_{s=\lambda(y_i)} \right\rangle \right. \\ & \left. + \left\langle (g(\lambda(y_i)) - y_i), \frac{d}{dz_k} \frac{d}{ds} g(s) \Big|_{s=\lambda(y_i)} \right\rangle \right]. \end{aligned} \tag{5}$$

This has a similar form as the first variation of $q(\lambda, Y)^2$ and for λ orthogonal to g $\hat{q}(\lambda, Y)$ is also minimal. However, there may exist additional local minima for which λ is not an orthogonal projection onto g.

Because this is a local optimization, one must consider the initialization. The parameters Z can be initialized by principal components, a spectral manifold learning method (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003), or by some heuristic that is particular to the application. The algorithm generalizes in a straightforward manner to m-dimensional manifolds by initializing Z to points in $\mathbf{R}^m$ and replacing dot products with the appropriate matrix-vector multiplications.

3.1 Model Complexity through Automatic Bandwidth Selection

Gerber et al. (2009) reported good results with cross-validation for minimizing the mean-squared projection distance $\hat{d}(\lambda, Y)^2$ starting from a λ that results in a smooth estimate g. However, the optimization was based on fixed bandwidths for the kernel regression and radial basis function expansion and thus imposed implicitly a regularization on the estimate.

In the proposed computational scheme, there are two bandwidth selection problems σ_λ for λ and σ_g for g that control model complexity. The bandwidth for λ determines a degree of smoothness in the coordinate mapping and thereby provides a form of regularization. For sufficiently small λ, the regularization effect is negligible but can negatively affect generalization performance. The bandwidth of g determines the smoothness of the conditional expectation estimate. By shrinking σ_g towards zero the estimate will be forced to pass through the training data.

For automatically deducing model complexity the bandwidth selection is incorporated into the optimization. The optimization alternates between gradient descent steps of the spatial locations Z (5) of the radial basis function expansion λ and gradient descent steps of the bandwidths

$$\frac{d}{d\sigma}\hat{q}(\lambda, Y)^2 = \left(\frac{\partial}{\partial\sigma_g}\hat{q}(\lambda, Y)^2, \frac{\partial}{\partial\sigma_\lambda}\hat{q}(\lambda, Y)^2\right)$$

with stopping based on held out data.

The empirical results bear out the theoretical arguments in Section 2: cross-validation for bandwidth selection for g and λ is not feasible when minimizing projection distance. The minimization of $\hat{d}(\lambda, Y)^2$ with bandwidth selection tends towards curves with increased curvature, possibly moving away from a *close by* principal curve that is smoother than the initial curve (see Figures 3 and 5). Even starting from smooth curves cross-validation is not a reliable indicator for estimating a good, approximating principal curve (see Figure 4). This, empirically, demonstrates the saddle point nature of the critical points. Minimizing $\hat{q}(\lambda, Y)$ mitigates these issues and shows a desirable objective-function landscape (see Figures 3, 4 and 5).

For minimizing $\hat{q}(\lambda, Y)^2$ stopping based on cross-validation data was unnecessary in all numerical experiments—the test error follows closely in step with the training error. This nice property is not unexpected and hinges directly on the theoretical results of the new formulation to principal curves. Over-fitting in the case of minimizing $\hat{q}(\lambda, Y)^2$ requires a λ which is close to orthogonal for training data but not on the test data. However, the objective function $\hat{q}(\lambda, Y)^2$ indicates no preference towards such a particular orthogonal λ on the training data and does not express a desire to increase the length/complexity of the curve in order to achieve a minima. Furthermore, a minima of $q(\lambda, Y)^2$ given the complete density p is also a minima for the empirical objective $\hat{q}(\lambda, Y)^2$ on a sample of Y. This is not true for the objective function $d(\lambda, Y)^2$ or even for least squares regression in the supervised setting (however, for regression cross-validation is a viable option as mentioned in the discussion in Section 1). These two observations provide an explanation of the remarkable property of the proposed estimation scheme that a regularization-free minimum on the training data is typically a desirable principal curve.

3.2 Computational Considerations

The observation on the properties of the regularization suggest the optimization strategy of Z, σ_λ and σ_g in Algorithm 1, which proves to be effective in practice (see Section 4).

```
Data: $Y$
Result: Conditional expectation curve
Initialize $Z$ from $Y$ ( isomap, PCA, LLE, ... ) ;
Initialize $\sigma_\lambda$ to average $k$-nearest neighbor distance of $Y$;
Initialize $\sigma_g$ to average $k$-nearest neighbor distance of $\lambda(Y)$ ;
while $\hat{q}(\lambda, Y)$ is improving do
    while $\hat{q}(\lambda, Y)$ is improving do
        Gradient descent step in $\sigma_\lambda$ ;
    end
    Gradient descent step in $\sigma_g$ ;
    Gradient descent step in $Z$ ;
end
```

Algorithm 1: Conditional expectation curve optimization strategy

Various alternatives and extensions for the algorithm proposed here are possible, such as the choice of the conditional expectation estimator and the modeling of λ. A naive implementation of the algorithm has a complexity of $O(n^3md)$ per gradient descent step with n the number of points, m the intrinsic dimensionality of the manifold and d the ambient dimensionality. This can be reduced to $O(nc^2md)$ by taking advantage of local nature of the kernel based formulation for both g an λ. For any y_i and z_i only approximately c data points, with c depending on the bandwidth parameter, are effectively contributing to the mapping. A detailed analysis of these run time complexity results is in Gerber et al. (2009). For further speed up one can consider reducing the number of parameters Z in the mapping λ. The mapping complexity of λ only depends on the complexity of the principal curve and can be independent of the number of data points. Since the complexity can conceivably vary spatially, one should consider an adaptive algorithm that introduces more or less *control points* Z in λ as required and varies the bandwidth spatially as well. While the spatial varying bandwidth is essentially included in the optimization through spreading the parameters Z differently, an implementation incorporating variable bandwidths is potentially more effective in practice.

4. Numerical Experiments

The numerical experiments illustrate the theoretical results on the critical point properties of $q(\lambda, Y)^2$ and $d(\lambda, Y)^2$ using the estimation strategy proposed in Section 3. Gerber et al. (2009) report extensive results on the optimization of $\hat{d}(\lambda, Y)^2$ with equal or improved performance compared to several principal curve algorithms in various settings. Thus, the comparison is representative to a typical principal curve estimation scheme based on minimizing the squared projection distance. The results demonstrate that automatic bandwidth selection leads to good results for $\hat{q}(\lambda, Y)^2$, while minimizing $\hat{d}(\lambda, Y)^2$ shows the expected behavior of tending towards curves with high curvature that pass close to the training data, even with early stopping based on held-out data.

4.1 Conditional Expectation Curves in the Plane

Here we demonstrate the observations on optimization and bandwidth selection on a 2D data set generated by $a(\phi) = (\sin(\phi), \cos(\phi))^t(1+\varepsilon)$. A 180 degree arc of radius one and $\varepsilon \approx \mathcal{N}(0, \eta^2)$ normal distributed noise added orthogonal to the arc. To generate n data points ϕ is sampled uniformly on $[0, \pi]$.

Figure 3 shows optimization results for $n = 150$ and $\eta = 0.15$ but with $Z = \sin(\phi)$ initialized close to a principal component. Starting from bandwidths $\sigma_g = 0.2$ and $\sigma_\lambda = 0.1$, the minimization of $\hat{d}(\lambda, Y)$ moves away from the principal component while $\hat{q}(\lambda, Y)$ approximately recovers the principal component. These observations also hold for other bandwidth initializations.

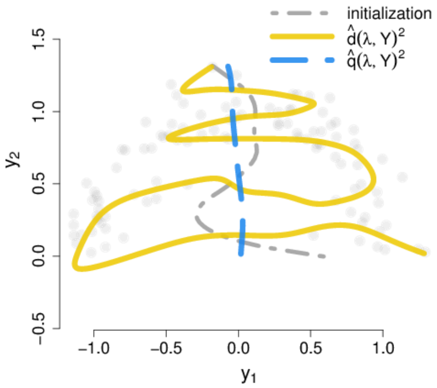

Figure 3: Minimization of $\hat{d}(\lambda, Y)^2$ and $\hat{q}(\lambda, Y)^2$ starting from a curve close to a principal component with bandwidths initialized to $\sigma_g = 0.2$ and $\sigma_\lambda = 0.1$.

Figures 4 and 5 show optimization results for $n = 150$, $\eta = 0.25$ and $Z = \phi$ initialized to the ground truth starting from different bandwidth initialization. Once again, minimizing $\hat{d}(\lambda, Y)^2$ favors curves with high curvature. The minimization of $\hat{q}(\lambda, Y)^2$ moves towards the ground truth, even for a small initial bandwidth.

The theoretical results give no guarantee/bounds on the radius of convergence around solutions. However,in all numerical experiments the optimization moved towards the anticipated solution.

4.2 Conditional Expectation Surfaces in 3D

The numerical results on algorithm behavior generalize to conditional expectation surfaces. This section replicates the results from Section 4.1 on the synthetic *Swissroll* data, a 2D-manifold in 3D-space (illustrated in Figures 6 and 7), commonly used to validate manifold learning techniques. The Swissroll is a parametric surface $s(h, \phi) = (\cos(\phi)\phi, h, \sin(\phi)\phi)$. The data for the experiments are generated by drawing observations $(h_i, \phi_i) \in H \times \Phi$, with $H = [0, 4]$ and $\Phi = [\pi, 3\pi]$ and adding normal $\mathcal{N}(0, \frac{1}{2})$ distributed noise orthogonal to $s(h_i, \phi_i)$. Note, this induces a nonuniform sample of the Swissroll, where the density in the ambient space decreases with increasing radius.

Figure 6 shows results for minimizing $\hat{d}(\lambda, Y)^2$ and $\hat{q}(\lambda, Y)^2$ on a training sample with $n = 1000$ drawn uniformly on $H \times \Phi$. The optimization of Z is initialized by a 2D isomap embedding using 15-nearest neighbors. Bandwidths are initialized by the average 75-nearest neighbor distance of $\lambda(Y)$ for σ_g and the average 25-nearest neighbor distance of Y for σ_λ. The same behavior as for the

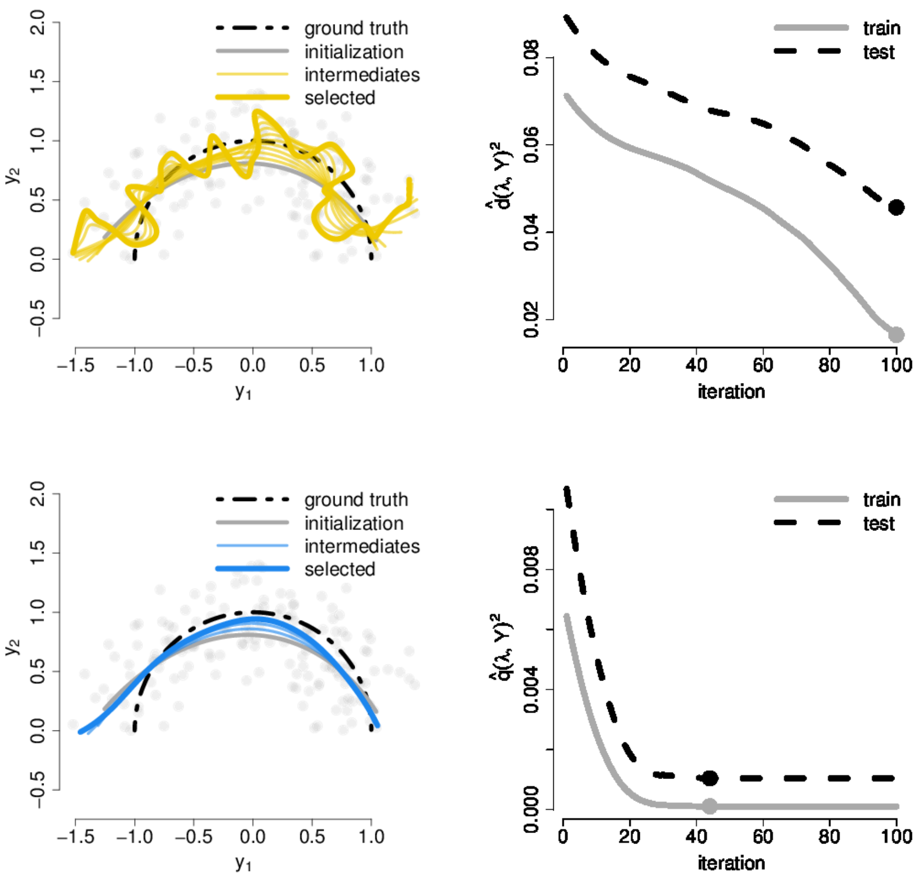

Figure 4: Minimization of (top) $\hat{d}(\lambda,Y)^2$ and (bottom) $\hat{q}(\lambda,Y)^2$ with automatic bandwidth selection starting from $\sigma_g = 1$ and $\sigma_\lambda = 0.1$. (left) Fitted curve with optimization path and (right) train and test error with points indicating minimal train and test error, respectively.

1D case in Section 4.1 holds. To qualitatively inspect the optimization, test data is created using points from a regular grid on $H \times \Phi$, again with normal $\mathcal{N}(0, \frac{1}{2})$ noise orthogonal to the surface added at each grid location.

Figure 7 shows optimization results starting from a surface close to a principal plane. Bandwidths are initialized by the average 75-nearest neighbor distance of $\lambda(Y)$ for σ_g and the average 25-nearest neighbor distance of Y for σ_λ. As in the 1D case $\hat{q}(\lambda,Y)^2$ moves towards the principal plane, while $\hat{d}(\lambda,Y)^2$ results in a highly curved fit that begins to fill the space.

4.3 Conditional Expectation Manifolds in Higher Dimensions

The final experiment reports results of fitting a 3D manifold on an image data set. The data set consists of 1965 images of a face with different facial expression with varying orientation. The images are 20×28 pixels with intensities in $[0,255]$ and treated as points in a 560-dimensional Euclidean space. To test the quality of automatic bandwidth selection normally-distributed ($\mathcal{N}(0,20)$) noise is added to each pixel. The data set is randomly split into 1000 training and 965 testing images. Then

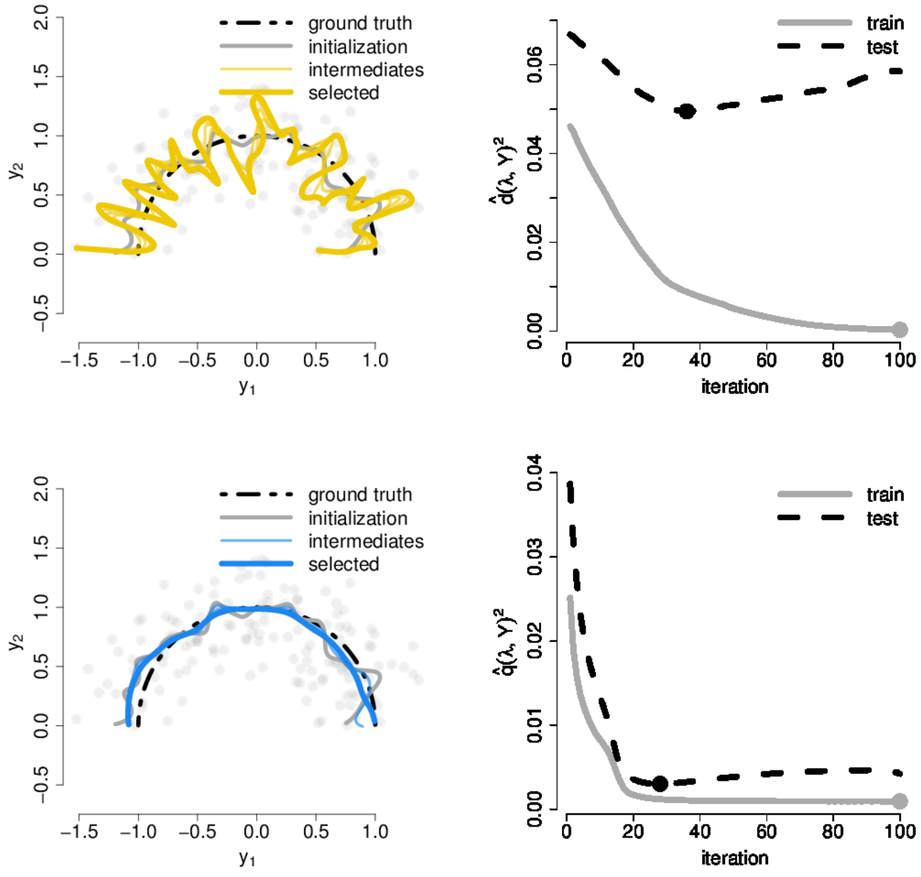

Figure 5: Minimization of (top) $\hat{d}(\lambda,Y)^2$ and (bottom) $\hat{q}(\lambda,Y)^2$ with automatic bandwidth selection starting from $\sigma_g = 0.1$ and $\sigma_\lambda = 0.1$. (left) Fitted curve with optimization path and (right) train and test error with points indicating minimal train and test error, respectively.

$\hat{q}^2(\lambda,Y)$ and $\hat{d}^2(\lambda,Y)$ is optimized on the training data with Z initialized to a 3D isomap embedding. The optimization is stopped based on the objective function value of the test data. Figure 8 shows the original and noisy images and the projection on the conditional expectation manifold for each optimization. While the mean squared projection distance is similar for both solutions, the optimization $\hat{q}^2(\lambda,Y)$ leads to qualitatively much better results. This underscores the observations in the introduction: mean squared error in itself is not a good indicator of model quality for nonparametric manifold estimation.

5. Conclusion

The conditional expectation curve formulation together with the introduction of a new objective function solves the model complexity problem in principal curve estimation. The resulting algorithm for finite data has the remarkable property that a regularization-free minimum on training data is typically a desired solution. The proposed formulation appears to transform the ill-posed

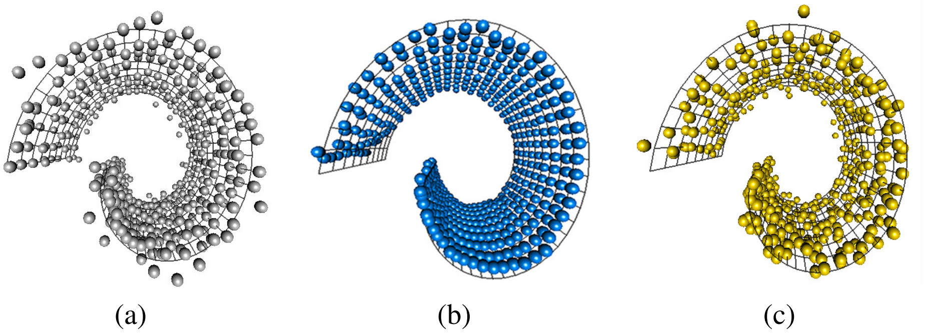

Figure 6: Optimization results on the Swissroll (wire frame). Projection $g(\lambda(y))$ of (a) test data after minimization (on a separate training sample shown in Figure 7) of (b) $\hat{q}(\lambda, Y)^2$ and (c) $\hat{d}(\lambda, Y)^2$. A good estimate projects the noisy test data points, located orthogonal to the wire frame, close to the intersections.

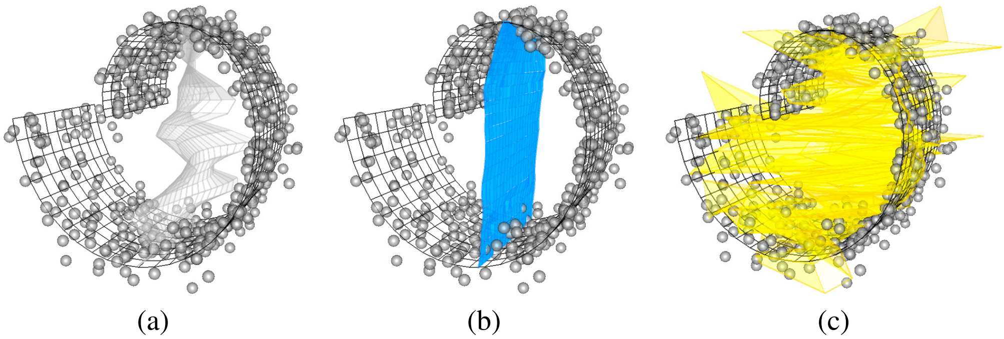

Figure 7: Minimization of (b) $\hat{q}(\lambda, Y)^2$ and (c) $\hat{d}(\lambda, Y)^2$ starting from an (a) initial surface close to a principal plane.

problem of principal curve estimation into a well-posed, or at least stable, problem. A formal investigation of the stability as defined in the framework of Mukherjee et al. (2006) could provide insight towards possible adaptations to other ill-posed problems.

Several important questions regarding the interplay between noise, curvature and sample size remain open. Of particular interest are the convergence rates of $\hat{q}(\lambda, Y)$ to $q(\lambda, Y)$ or error bounds on the corresponding conditional expectation curves. An analysis of the requirements imposed by projection distance, curvature and sample size on σ_λ and implications on alternative formulations for modelling λ is an interesting direction for future research.

Acknowledgments

The authors thank Thomas Fletcher for the helpful discussions on Lemma 2. This work was supported by the NSF grant CCF-073222.

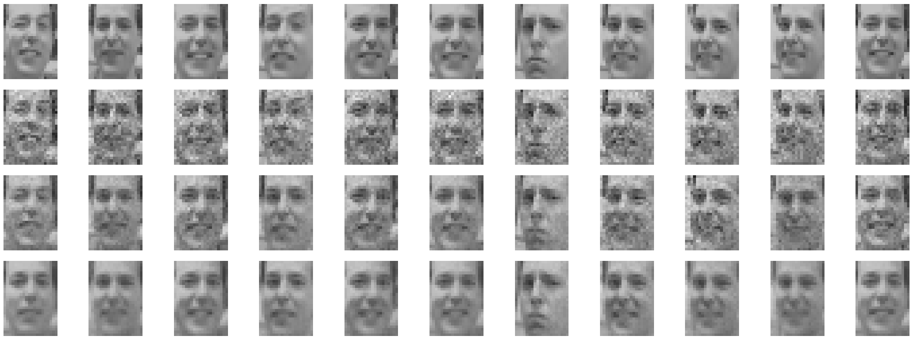

Figure 8: Optimization results on (1st row) face data set. Test data (2nd row) with $\mathcal{N}(0,20)$ noise added, projected onto the conditional expectation manifold after minimization of (3rd row) $\hat{d}(\lambda,Y)^2$ and (4th row) $\hat{q}(\lambda,Y)^2$ on training data also with $\mathcal{N}(0,20)$ noise added.

References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

G. Biau and A. Fischer. Parameter selection for principal curves. *IEEE Transactions On Information Theory*, 58(3):1924 –1939, 2012.

K-Y. Chang and J. Ghosh. A unified model for probabilistic principal surfaces. *IEEE Transaction On Pattern Analysis And Machine Intelligence*, 23(1):22–41, 2001.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

T. Duchamp and W. Stuetzle. Extremal properties of principal curves in the plane. *The Annals of Statistics*, 24(4):1511–1520, 1996.

H. Federer. *Geometric Measure Theory*. Springer, NY, 1969.

H. Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6): 615–627, 1973.

S. Gerber. *Cems: Conditional Expectation Manifolds*, 2011. URL `http://CRAN.R-project.org/package=cems`. R package version 0.1.

S. Gerber, T. Tasdizen, and R. Whitaker. Dimensionality reduction and principal surfaces via kernel map manifolds. In *IEEE 12th International Conference on Computer Vision*, pages 529–536, 2009.

W. Härdle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13(4), 1985.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 (406):502–516, 1989.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006.

B. Kégl, A. Krzyzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transaction On Pattern Analysis Machine Intelligence*, 22(3):281–297, 2000.

T. Kohonen. *Self-organized Formation of Topologically Correct Feature Maps*, pages 509–521. MIT Press, Cambridge, MA, USA, 1988.

P. Meinicke, S. Klanke, R. Memisevic, and H. Ritter. Principal surfaces from unsupervised kernel regression. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 27(9):1379–1391, 2005. ISSN 0162-8828.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(550), 2000.

A. J. Smola, S. Mika, B. Schölkopf, and R. C. Williamson. Regularized principal manifolds. *Journal Of Machine Learning Research*, 1:179–209, 2001.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(550):2319–2323, 2000.

R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

H. Wang and T. C. M. Lee. Automatic parameter selection for a k-segments algorithm for computing principal curves. *Pattern Recognition Letteerss*, 27(10):1142–1150, 2006.

G. Watson. Smooth regression analysis. *Sankhya, Series*, A(26):359–372, 1964.