1. **Statistical Shape Analysis (SSA)**

**Shape of an object:** all information of the object that is invariant with respect to similarity transformations on Euclidean space (rotations, translations, dilations). Data is a 2D or 3D cloud point.

**Shape analysis of manufactured parts**

An object is described by a \( k \times n \) configuration matrix \( X \) (\( m = 2 \) or 3, \( k \) could be very large).

- Assumed model in SSA: \( n \) measured objects \( X_i = \beta_i (\mu + E_i) \Gamma_i + 1 \gamma_i \), \( \vee \Sigma \sim N(0, \Sigma) \).
- Generalized Procrustes Analysis (GPA): a method for estimating the mean shape \( \mu \) from a sample of \( n \) objects that may have different scales, orientations and locations.

**GPA solutions**

\[
G(X_1, X_2, ..., X_n) = \min_{\beta_i, \Gamma, \gamma_i} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \left( \beta_i X_i \Gamma_j + 1 \gamma_i - (\beta_i X_i \Gamma_j + 1 \gamma_i) \right)^2
\]

- Estimate: \( \hat{X}_i = \hat{\beta}_i \hat{X}_i \hat{\Gamma}_i + 1 \hat{\gamma}_i \), \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i \) and \( \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T \).

**Analysis of experiments with shape responses: Two-way ANOVA for shapes.**

- Two-way ANOVA model for observed objects: \( E(X_{ij}) = \mu + \tau_i + \beta_j + (\tau \beta)_{ij} \). Define \( d^2_{ij}(X_{ij}, X_{jk}) = G(X_{ij}, X_{jk}) \) (procrustes distance: a metric in the non-euclidean shape space manifold).

**MANOVA cannot be used since usually \( k \times m > h(n-1) \).**

- Test: \( H_0^{(1)} : \tau_i = 0, H_0^{(2)} : \beta_j = 0 \) and \( H_0^{(3)} : (\tau \beta)_{ij} = 0 \) \( SS_{total} = SS_A + SS_B + SS_{AB} + SS_{error} \) where \( SS_{total} = \sum_{i=1}^{n} \sum_{j=1}^{k} d^2(X_{ij}, X_{jk}) \), \( SS_A = bn \sum_{i=1}^{n} \sum_{j=1}^{k} d^2(X_{ij}, X_{kj}) \), \( SS_B = an \sum_{i=1}^{n} \sum_{j=1}^{k} d^2(X_{ij}, X_{kj}) \), \( SS_{AB} = \sum_{i=1}^{n} \sum_{j=1}^{k} d^2(X_{ij}, X_{kj}) \), \( SS_{error} = \sum_{i=1}^{n} \sum_{j=1}^{k} d^2(X_{ij}, X_{kj}) \).

- \( F_{0}^{0} = MS_A/MS_{error}, \) etc., distribution results hold under isotropic variance if the variance is small (shapes are “close”).

**Normal isotropic assumption probably unrealistic:** use two-way Permutation ANOVA for Shapes (Del Castillo and Colosimo, 2011). More powerful than other tests for shape effect detection. Multiple comparisons based on the procrustes metric derived.

- Usually effective estimators are pre-shapes (i.e., normalized, small differences hard to visualize: \( \hat{\tau}_i = X_{ij} - X_{jk}, \hat{\beta}_j = X_{ij} - X_{jk} \)).
- Permutation field (“quiver”: \( q \)) plots relating the effects to the main shape \( \hat{\mu} = \hat{X}_{...} \) (simulated responses)

![Figure 1. Main effect on the shape for factor A (left) and B (right).](image)

2. **Gaussian Process (GP) Modeling**

Let \( Y(x) \in \mathbb{R}^d \) be a stochastic process where \( D \) is a fixed subset of \( r \)-dimensional Euclidean space. If every finite vector \( Y(\mathbf{x}_1), Y(\mathbf{x}_2), ..., Y(\mathbf{x}_n) \) for \( n \geq 1 \) has a multivariate normal distribution, the process is said to be a **Gaussian Process**.

**Geodesic GPs for reconstructing free-form manufactured parts**

- Consider measurement of a free form surface. Dataset is a 3D unstructured cloud point \((x, y, z)\) data. Reconstruct the **true surface** for: **Inspection**, “Reverse engineering” or **Statistical Process Control**. Tasks easier if a model of the true surface available.

- Usually approach: model is \( z(x, y) \rightleftharpoons \text{unclear why } z \) is the “response” and \((x, y)\) the “locations”. Assumes variables correlated as a function of the euclidean distance between their locations in the XY plane. **BUT** the \((x, y, z)\) data are on a 2D manifold, not on any plane.

**GGM model:** use a GP for each coordinate surface (parametric surface, compatible with CAD). Correlation between points assumed over geodesic distances on the non-euclidean surface. Requires an “as-isometric-as-possible” parameterization, i.e., a mapping \( p : D \subset \mathbb{R}^2 \rightarrow S \subset \mathbb{R}^3 \) that preserves (geodesic) distances. Parametrization problem studied in **computer graphics**.

![Figure 2. Gaussian Process with geodesic mapping](image)

**True underlying surface \( S \) can be observed only with error:** \( \mathbf{m}(w) = (m_x(w), m_y(w), m_z(w))' = p(w) + \varepsilon(w), \mathbf{w} = (x, y, v) \in D \) where \( \varepsilon(w) \sim N(0, \Sigma) \). Given a parameterization, true surface modeled with a smooth spatial GP (the “state” equation):

\[
p(w) = (z(u, v), y(u, v), z(u, v))' = (\beta_1 f_1(w), \beta_2 f_2(w), \beta_3 f_3(w))' + \delta(w), \mathbf{w} = (u, v) \in D
\]

where \( \delta(w) \) zero-mean, smooth (no-nugget), 3-dimensional vector stationary Gaussian covariance functions \( c_h, c_y, c_z \), respectively, where \( h = w - w_o \). Predictions are:

\[
p((u_0, v_0)) = f((u_0, v_0))' + c_y \Sigma_y^{-1} (M_0 - F_0 \hat{\beta}_0), \quad \{x, y, z\}
\]

Note: \( c_y \) does not contain the nugget; we predict (reconstruct) \( p(w_0, v_0) \) not the observed \( m(w_0, v_0) \).
### Geodesic Gaussian Process for free-form surface reconstruction (cont.)

**Manifold learning and computer graphics** algorithms tested for finding a (n, ε) parametrization. Ideal parametrization: an isometry ($ρ = 1$). GGP prediction errors improved one of order of magnitude over euclidean GP due to better modeling of curved features.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference(s)</th>
<th>$ρ$ estimate</th>
<th>MSE with error</th>
<th>MSE without error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESMC</td>
<td>Levy et al., 2008</td>
<td>0.9976</td>
<td>0.0092</td>
<td>0.0084</td>
</tr>
<tr>
<td>GMAP</td>
<td>Little et al., 2011</td>
<td>0.0076</td>
<td>0.0076</td>
<td></td>
</tr>
<tr>
<td>GLEL</td>
<td>Rossel et al., 2005</td>
<td>0.9992</td>
<td>0.0092</td>
<td></td>
</tr>
<tr>
<td>KGPA</td>
<td>Donoso et al., 2003</td>
<td>0.0076</td>
<td>0.0076</td>
<td></td>
</tr>
<tr>
<td>KPCA</td>
<td>Stoiklieva et al., 1998</td>
<td>0.0076</td>
<td>0.0076</td>
<td></td>
</tr>
<tr>
<td>Isomap</td>
<td>Tennenbaum, 2000</td>
<td>0.0095</td>
<td>0.0095</td>
<td></td>
</tr>
</tbody>
</table>

**3. Confidence regions of response surface optima for product formulation**

We wish to find a confidence region (CR) for the function:

$$h(x; \hat{\beta}) = \max_{(x, \beta)} f(x, \beta), \quad x \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^m$$

where $f(x, \beta)$ is either a parametric regression model in $x$ or a nonparametric Thin Plate Spline model in $x$ fitted from a sample of noisy observations $y = f(x, \beta) + \epsilon$. The solution $x^*$ is only a *point estimate* on the true optimum.

**Motivation.** The shape and location of the CR provides alternative optimal formulation settings. Spline models are widely used in engineering to locate best regions of operation of a process, but no methodology exists for obtaining a CR on $x^*$. Such a CR answers the therapeutic suspension problem found in the pharmaceutical industry, and, by extension, complicated product formulation problems (chemical and food industries). How to compute a CR? Particular interest is in non-normal datasets.

**CS** (confidence set) method for finding confidence regions of functions of parameters. The method is based on the following steps:

1. **Obtain a 100(1-α)% CR for $\hat{\beta}$ from the asymptotic distribution of $\hat{\beta}$.**
2. **For each $\beta \in \mathbb{R}^m$, evaluate $h(\beta)$.**
3. **Let $\mathcal{C}_{\beta} = \{x \in \mathbb{R}^n | h(x, \beta) = h(\beta)\}$ for all $\beta \in \mathbb{R}^m$.**

To estimate this confidence region, we propose **bootstrapping** in steps 1 and 3:

1. **Obtain an estimate of the 100(1-α)% CR for $\beta$ by bootstrapping $B$ instances of $\hat{\beta}$.** These instances make $\mathcal{C}_{\beta}^B$.
2. **For each $\beta \in \mathbb{C}_{\beta}^B$, evaluate $h(\beta)$.**
3. **Let $\mathcal{C}_{\beta}^B = \{x \in \mathbb{R}^n | h(x, \beta) = h(\beta)\}$ for all $\beta \in \mathbb{C}_{\beta}^B$.**

### Data depth bootstrapping CR methods

In nonparametric models the number of parameters $\beta$ is by definition infinite. In the case of Splines, however, even though the model fitting is an optimization over an infinite-dimensional Hilbert space $H$:

$$\hat{f} = \arg\min_{f \in H} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda ||P||^2 \right\}$$

(where $\lambda > 0$ trade-offs smoothness vs. MSE) the remarkable *Kimeldorf- Wahba* theorem indicates that the solution $\hat{f}$ is given by a finite dimensional operation that depends on a finite number of parameters $\beta$:

$$\hat{f} = \sum_{i=1}^B \delta_i \hat{\phi}_i(x) + \sum_{i=1}^B \xi_i \hat{\phi}_i(x)$$

so let $\beta = (\delta', \xi')$ in the bootstrapping algorithm. This may result in high dimensional vectors of parameters. Need methods to construct high dimensional confidence regions for the parameters of a linear model.

We use a **data depth** measure of the centrality of a point with respect to the rest of the data. Given a set of points $F = \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{R}^n$, the data depth measure of an additional point $x$ is a real-valued function $d(x|F)$. Many such functions exist; Tukey’s data depth is:

$$D_F(x, F) = \min_{\|\omega\|_0=1} \text{card}(\{x' \in F | x' \leq x\})$$

In the CS-bootstrapping method applied to problem (1), we order the $B$ instances $\hat{\beta}$ according to $D_F(x, F)$ and trim the α% outermost (the α% with lowest $D_F$ value). This yields $\mathcal{C}_{\beta}^B$ in step 1.

**An example in Evolutionary Biology.** Theory predicts that when a population is subject to stabilizing selection over time the population mean should evolve to the peak of the fitness surface. Experiments in mice and insects vary the components of diets (carbs and P) and measure responses that are surrogates of fitness (e.g., no. of eggs placed by a female insect). Of practical importance for humans are lifetime experiments with diets. These are *mixture-amout* experiments.

**Some selected journal papers from this work**


**Acknowledgements**

1. National Science Foundation grants DMI 0826609, DMI 0958336, CMMI 0830414, CMEE 0827857 and CMEE 1537887.
2. Ministero dell’Istruzione, dell’Università e della Ricerca (Italy) grant FISR-08PFRE1709-000.
3. Prof. Bianca Coloma, Dept. of Production Engineering, Politecnico di Milano, Italy.
4. Dr. John Hunt, Director of Statistics, GlaxoSmithKline.