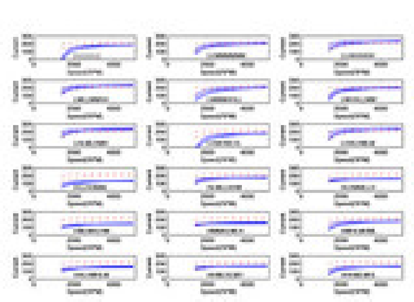
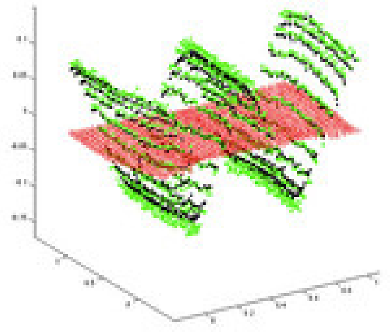




Some Research Themes in Engineering Statistics and Machine Learning

P.I.: Enrique del Castillo, ESAMLab (Engineering Statistics and Machine Learning Laboratory)
Dept. of Industrial & Manufacturing Engineering and
Department of Statistics, PSU

PennState College of Engineering
HAROLD AND INGE MARCUS DEPARTMENT OF INDUSTRIAL AND MANUFACTURING ENGINEERING

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 and dilations). Data is a 2D or 3D cloud point.

Shape analysis of manufactured parts

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

- Assumed model in SSA: n measured objects $\boldsymbol{X}_i = \beta_i(\boldsymbol{\mu} + \boldsymbol{E}_i)\boldsymbol{\Gamma}_i + 1_k\boldsymbol{\gamma}_i^T$, $\quad vec(\boldsymbol{E}) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$
- Generalized Procrustes Analysis (GPA): a method for estimating the mean shape $\boldsymbol{\mu}$ from a sample of n objects that may have different scales, orientations and locations.
- GPA solves $\quad G(\boldsymbol{X}_1, \boldsymbol{X}_2, .., \boldsymbol{X}_n) = \min_{\beta_i, \boldsymbol{\Gamma}_i, \gamma_i} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} ||\beta_i \boldsymbol{X}_i \boldsymbol{\Gamma}_i + 1_k\gamma_i' - (\beta_j \boldsymbol{X}_j \boldsymbol{\Gamma}_j + 1_k\gamma_j')||^2$
- Estimate: $\boldsymbol{X}_i^p = \widehat{\beta}_i \boldsymbol{X}_i \widehat{\boldsymbol{\Gamma}}_i + \boldsymbol{1}_k \widehat{\boldsymbol{\gamma}}_i^T$, $\quad \widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^p \quad$ and $\quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}(\boldsymbol{X}_i^p - \widehat{\boldsymbol{\mu}})\mathrm{vec}(\boldsymbol{X}_i^p - \widehat{\boldsymbol{\mu}})'$.

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

- Two factor experiment. Model for observed objects has: $\quad E[\boldsymbol{X}_{ijl}] = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\beta}_j + (\boldsymbol{\tau\beta})_{ij}$, $\; i = 1, ..., a; j = 1, ..., b; l = 1, ..., n$. Define $d_F^2(\boldsymbol{X}_1, \boldsymbol{X}_2) = G(\boldsymbol{X}_1, \boldsymbol{X}_2)$ (**procrustes distance:** a metric in the non-euclidean shape space manifold). **Note:** MANOVA cannot be used since usually $k \cdot m > a \cdot b(n-1)$
- Test $H_0^{(1)} : \boldsymbol{\tau}_i = \boldsymbol{0}$, $H_0^{(2)} : \boldsymbol{\beta}_j = \boldsymbol{0}$ and $H_0^{(3)} : (\boldsymbol{\tau\beta})_{ij} = \boldsymbol{0}$: $\quad SS_{total} \approx SS_A + SS_B + SS_{AB} + SS_{error}$ where $SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} d_F^2(\boldsymbol{X}_{ijl}^p, \overline{\boldsymbol{X}}_{\bullet\bullet\bullet})$, $SS_A = bn\sum_{i=1}^{a} d_F^2(\overline{\boldsymbol{X}}_{i\bullet\bullet}, \overline{\boldsymbol{X}}_{\bullet\bullet\bullet})$, $SS_B = an\sum_{j=1}^{b} d_F^2(\overline{\boldsymbol{X}}_{\bullet j\bullet}, \overline{\boldsymbol{X}}_{\bullet\bullet\bullet})$, $SS_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b} d_F^2(\overline{\boldsymbol{X}}_{ij\bullet} - (\overline{\boldsymbol{X}}_{i\bullet\bullet} - \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}) - (\overline{\boldsymbol{X}}_{\bullet j\bullet} - \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}), \overline{\boldsymbol{X}}_{\bullet\bullet\bullet})$, $SS_{error} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n} d_F^2(\boldsymbol{X}_{ijl}^p, \overline{\boldsymbol{X}}_{ij\bullet})$.
- $F_0^{(1)} = \mathrm{MS}_A/\mathrm{MS}_{error}$, 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.
- Usual effect estimators are pre-shapes (i.e., normalized), small differences hard to **visualize**: $\widehat{\boldsymbol{\tau}}_i = \overline{\boldsymbol{X}}_{i\bullet\bullet} - \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}$, $\widehat{\boldsymbol{\beta}}_j = \overline{\boldsymbol{X}}_{\bullet j\bullet} - \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}$, $\widehat{(\boldsymbol{\tau\beta})}_{ij} = \overline{\boldsymbol{X}}_{ij\bullet} - \overline{\boldsymbol{X}}_{i\bullet\bullet} - \overline{\boldsymbol{X}}_{\bullet j\bullet} + \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}$
- Use **vector field** ("quiver") plots relating the effects to the mean shape $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}}_{\bullet\bullet\bullet}$ (simulated responses)

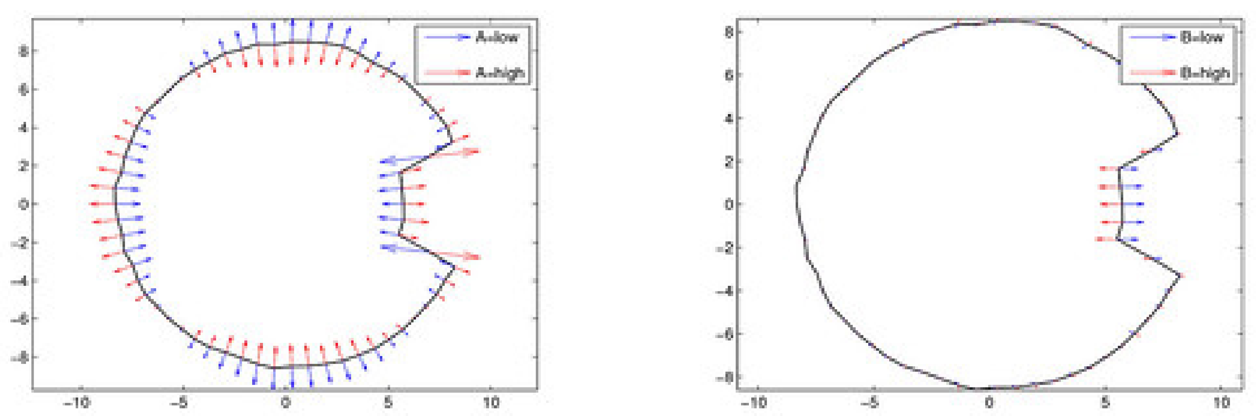


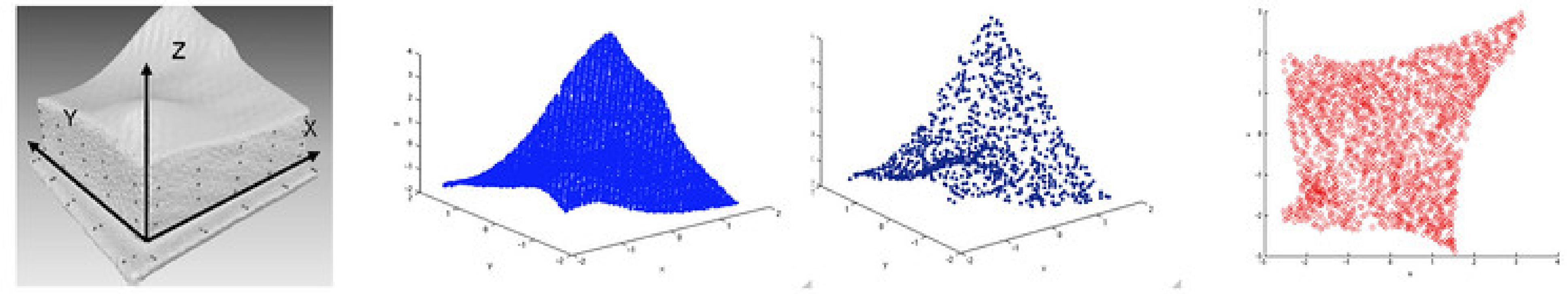
Figure 1. Main effect on the shape for factor A (left) and B (right).

2. Gaussian Process (GP) modeling

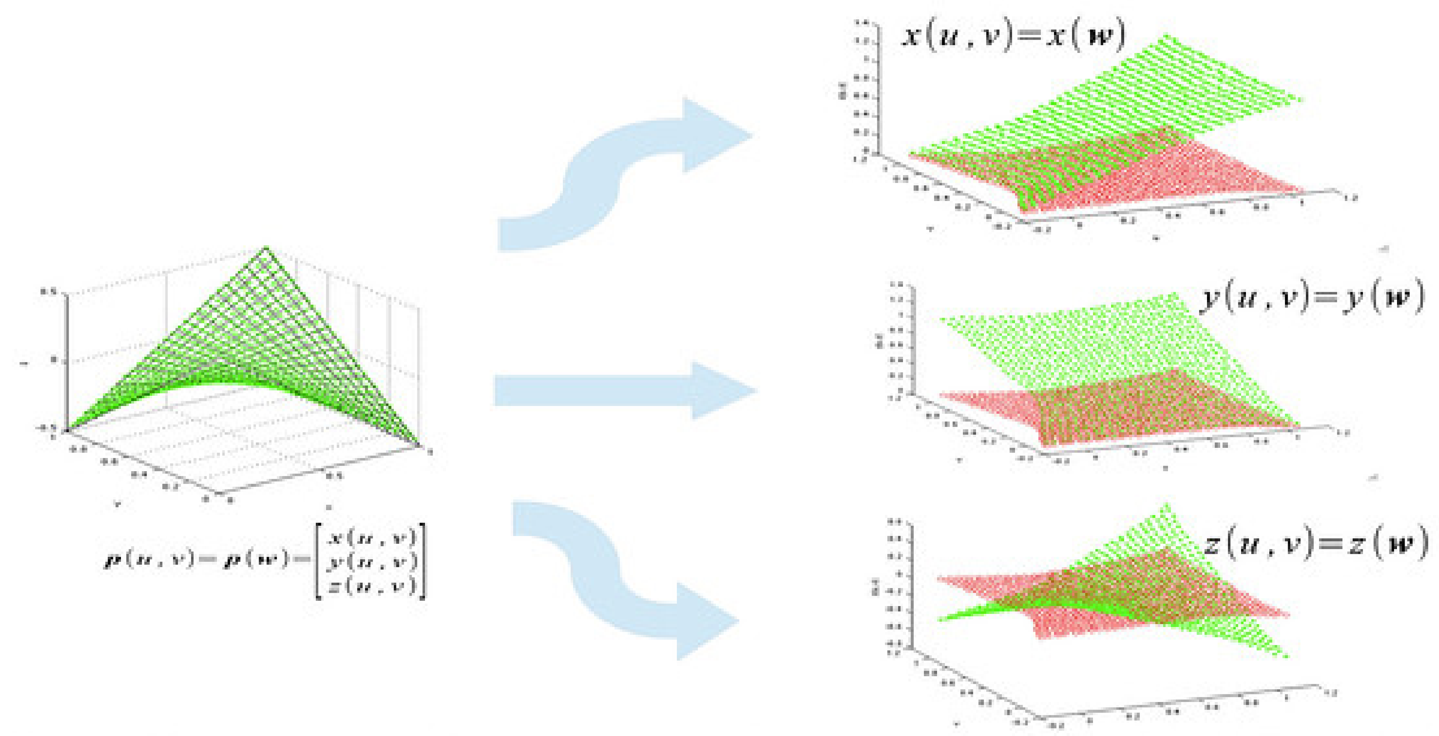
Let $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in D\}$ be a stochastic process where D is a fixed subset of r-dimensional **Euclidean** space. If every finite vector $\{Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$ for $n \geq 1$ has a multivariate normal distribution, the process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

Geodesic GP's for reconstructing free-form manufactured parts

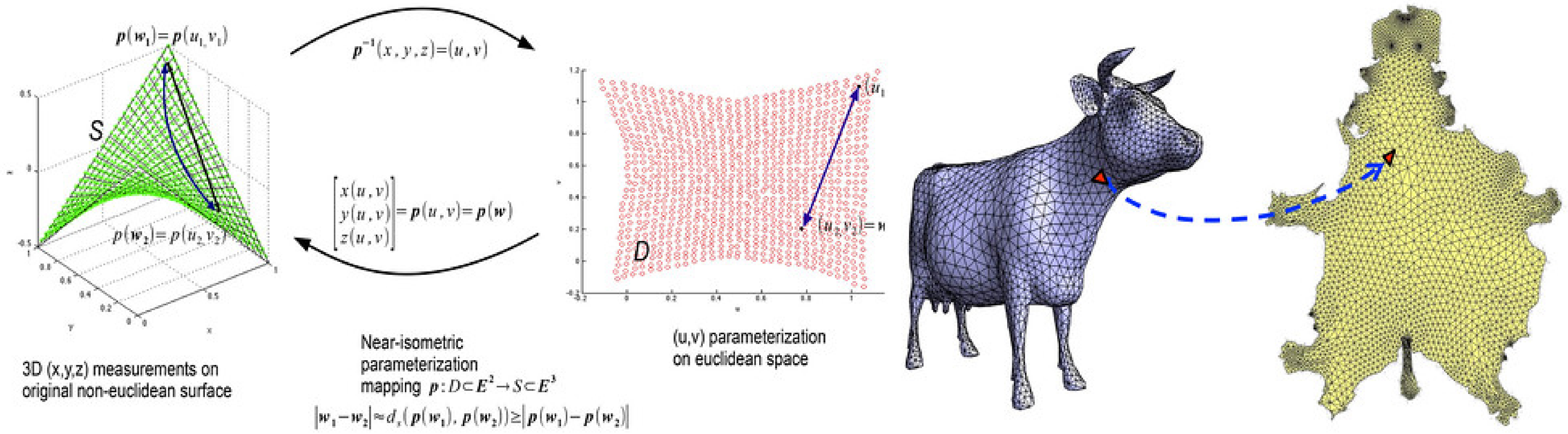
- Consider measurement of a *free form* surface. Dataset is a 3D unstructured **cloud point** ((x, y, z) data). Reconstruct the *true surface* for: **Inspection**, **"Reverse engineering"** or **Statistical Process Control**. Tasks easier if a **model** of the true surface available.
- **Usual approach**: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



- Typical form of a CAD model: patches of **parametric surfaces** (e.g., IGES standard) drawn using Non-Uniform Rational B-Splines (NURBS). A NURBS surface is a function $p : D \subset \mathbb{E}^2 \to S \subset \mathbb{E}^3$ of the form $\boldsymbol{p}(u, v) = (x(u, v), y(u, v), z(u, v))'$ where $(u, v) \in D$ are surface coordinates (or '**parameters**').
- NURBS very useful for drawing surfaces using CAD software, not so nice for fitting them from noisy data.
- A parameterized surface patch can be decomposed in its euclidean coordinate functions $x(u, v), y(u, v)$, and $z(u, v)$:



GGP 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 $p : D \subset \mathbb{E}^2 \to S \subset \mathbb{E}^3$ that preserves (geodesic) distances. Parametrization problem studied in **computer graphics**.



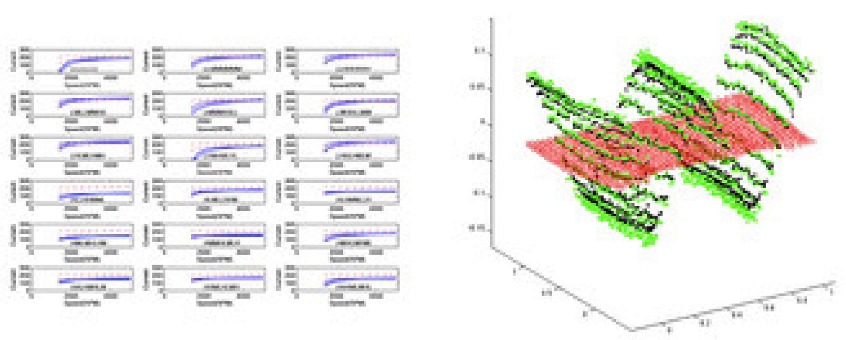
True underlying surface S can be observed only with error: $\boldsymbol{m}(\boldsymbol{w}) = (m_x(\boldsymbol{w}), m_y(\boldsymbol{w}), m_z(\boldsymbol{w}))' = \boldsymbol{p}(\boldsymbol{w}) + \boldsymbol{\varepsilon}(\boldsymbol{w})$, $\quad \boldsymbol{w} = (u, v) \in D$ where $\boldsymbol{\varepsilon}(\boldsymbol{w}) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$. Given a parametrization, true surface modeled with a smooth spatial GP (the "state" equation):

$$\boldsymbol{p}(\boldsymbol{w}) = (x(u, v), y(u, v), z(u, v))' = (\boldsymbol{\beta}_x' \boldsymbol{f}_x(\boldsymbol{w}), \boldsymbol{\beta}_y' \boldsymbol{f}_y(\boldsymbol{w}), \boldsymbol{\beta}_z' \boldsymbol{f}_z(\boldsymbol{w}))' + \boldsymbol{\delta}(\boldsymbol{w}), \quad \boldsymbol{w} = (u, v) \in D$$

where $\boldsymbol{\delta}(\boldsymbol{w})$ is a zero-mean, smooth (no-nugget), 3-dimensional vector stationary Gaussian Process with covariance functions $c_x(\boldsymbol{h})$, $c_y(\boldsymbol{h})$, and $c_z(\boldsymbol{h})$, respectively, where $\boldsymbol{h} = \boldsymbol{w}_i - \boldsymbol{w}_j$ Predictions are:

$$\hat{p}_\bullet(u_0, v_0) = \boldsymbol{f}(u_0, v_0)' \hat{\boldsymbol{\beta}}_\bullet + \boldsymbol{c}_{p_\bullet}' \boldsymbol{\Sigma}_\bullet^{-1}(\mathcal{M}_\bullet - \boldsymbol{F}_\bullet \hat{\boldsymbol{\beta}}_\bullet), \quad \bullet = \{x, y, z\}$$

Note: $c_{p_\bullet}$ *does not* contain the nuggets; **we predict (reconstruct)** $\boldsymbol{p}(u_0, v_0)$ not the observed $\boldsymbol{m}(u_0, v_0)$.



Some Research Themes in Engineering Statistics and Machine Learning (cont.)

P.I.: Enrique del Castillo*, ESAMLab (Engineering Statistics and Machine Learning Laboratory)
Dept. of Industrial & Manufacturing Engineering and
Department of Statistics, PSU

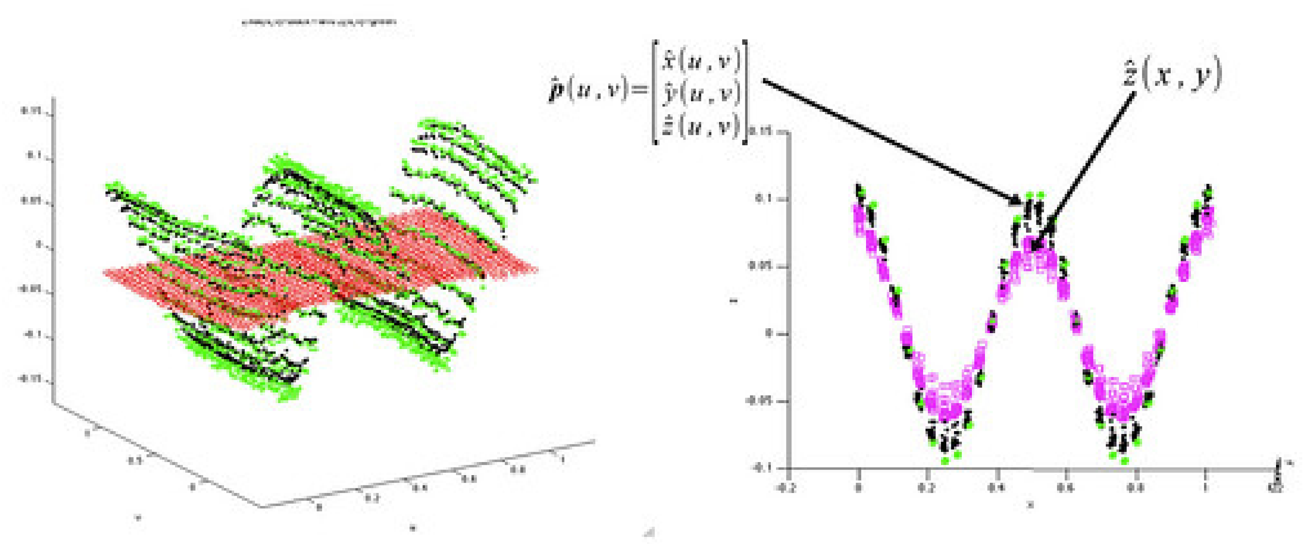
PennState College of Engineering
HAROLD AND INGE MARCUS DEPARTMENT OF INDUSTRIAL AND MANUFACTURING ENGINEERING

Geodesic Gaussian Process for free-form surface reconstruction (cont.)

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

Algorithm	Reference(s)	Estimated correlation ($\hat{\rho}$) No meas. error	Estimated correlation ($\hat{\rho}$) With meas. error
LSCM	Levy et al., 2008	0.9291	0.8784
ARAP	Liu et al., 2011	**0.9976**	**0.9953**
LLE	Roweis et al., 2000	0.9420	0.8998
HLLE	Donoho et al., 2005	0.9442	0.9434
KPCA	Shölkopf et al., 1998	0.9557	0.9557
Isomap	Tenenbaum et al., 2000	**0.9995**	**0.9984**

n	MSE_{3D}	MSP_{GGP}	$\text{MSP}_{z(x,y)}$
400	0.00761	**0.02435**	**0.04552**
	(0.00089)	(0.00693)	(0.00187)
900	0.00785	**0.01284**	**0.02946**
	(0.00132)	(0.00284)	(0.00148)
1600	0.00752	**0.00954**	**0.01957**
	(0.00099)	(0.00123)	(0.00112)

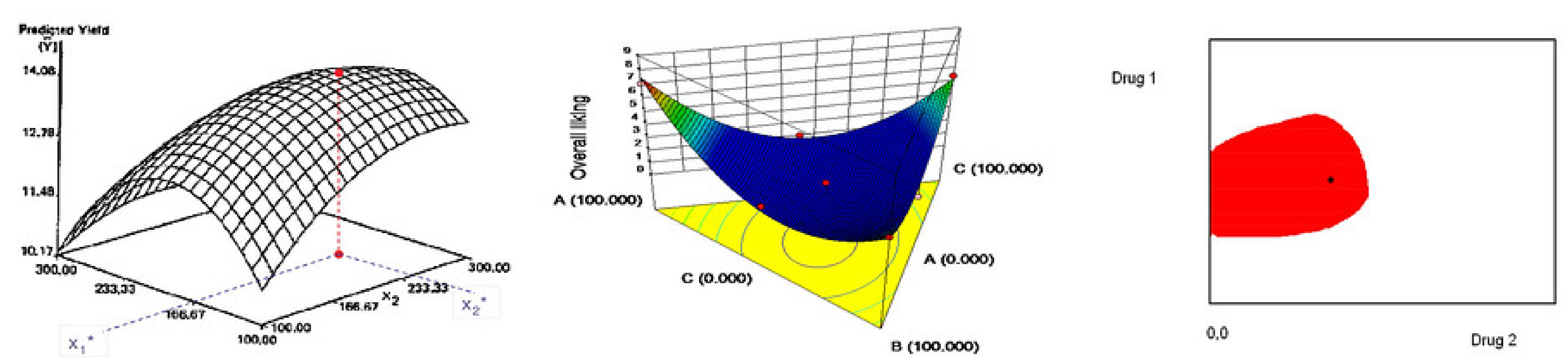


3. Confidence regions of response surface optima for product formulation

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

$$h(\mathbf{x}; \hat{\boldsymbol{\beta}}) = \arg\max f(\mathbf{x}, \hat{\boldsymbol{\beta}}), \quad \mathbf{x} \in \mathbb{R}^{\mathbf{k}}, \quad \boldsymbol{\beta} \in \mathbb{R}^{\mathbf{m}} \tag{1}$$

where $f(\mathbf{x}, \hat{\boldsymbol{\beta}})$ is *either* a parametric regression model in $\mathbf{x}$ *or* a nonparametric Thin Plate Spline model in $\mathbf{x}$ fitted from a sample of noisy observations $y = f(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$. The solution $\mathbf{x}^*$ 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 in science to locate best regions of operation of a process, but no methodology exists for obtaining a CR on $\mathbf{x}^*$. Such a CR answers the *therapeutic synergism 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 $100(1-\alpha)\%$ CR for $\boldsymbol{\beta}$ from the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

2 For each $\boldsymbol{\beta} \in \text{CR}^{\boldsymbol{\beta}}_{1-\alpha}$, evaluate $h(\boldsymbol{\beta})$.

3 Let $\text{CR}^{h(\boldsymbol{\beta})}_{1-\alpha} = \{\tau \in \mathbb{R}^k | \tau = h(\boldsymbol{\beta}) \text{ for all } \boldsymbol{\beta} \in \text{CR}^{\boldsymbol{\beta}}_{1-\alpha}\}$

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

$\mathbf{1_B}$ Obtain an estimate of the $100(1-\alpha)\%$ CR for $\boldsymbol{\beta}$ by bootstrapping B instances of $\hat{\boldsymbol{\beta}}$. These instances make $\widehat{\text{CR}}^{\boldsymbol{\beta}}_{1-\alpha}$;

$\mathbf{2_B}$ For each $\boldsymbol{\beta} \in \widehat{\text{CR}}^{\boldsymbol{\beta}}_{1-\alpha}$, evaluate $h(\boldsymbol{\beta})$.

$\mathbf{3_B}$ Let $\widehat{\text{CR}}^{h(\boldsymbol{\beta})}_{1-\alpha} = \{\tau \in \mathbb{R}^k | \tau = h(\boldsymbol{\beta}) \text{ for all } \boldsymbol{\beta} \in \widehat{\text{CR}}^{\boldsymbol{\beta}}_{1-\alpha}\}$

Data depth bootstrapping CR methods

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

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x_i}))^2 + \lambda ||\mathbf{Pf}||^2$$

(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 $\boldsymbol{\beta}$:

$$\hat{f} = \sum_{i=v}^{p} d_v \phi_v(\mathbf{x}) + \sum_{i=1}^{n} c_i \xi(\mathbf{x})$$

so let $\boldsymbol{\beta} = (\boldsymbol{d}', \boldsymbol{c}')'$ 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 $F = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n}\}$, $\mathbf{x_i} \in \mathbb{R}^{\mathbf{k}}$, the data depth measure of an additional point $\mathbf{x}$ is a real-valued function $d(\mathbf{x}|F)$. Many such functions exist; Tukey's data depth is:

$$D_T(\mathbf{x}, F) = \min_{||\mathbf{u}||=1} \text{card}\{\mathbf{u}'\mathbf{x_i} \leq \mathbf{u}'\mathbf{x}\}$$

In the CS-bootstrapping method applied to problem (1), we order the B instances $\widehat{\boldsymbol{\beta}}$ according to $D_T(\mathbf{x}, F)$ and trim the α % outermost (the α % with lowest D_T value). This yields $\widehat{\text{CR}}^{\boldsymbol{\beta}}_{1-\alpha}$ in step $\mathbf{1_B}$.

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-amount** experiments.

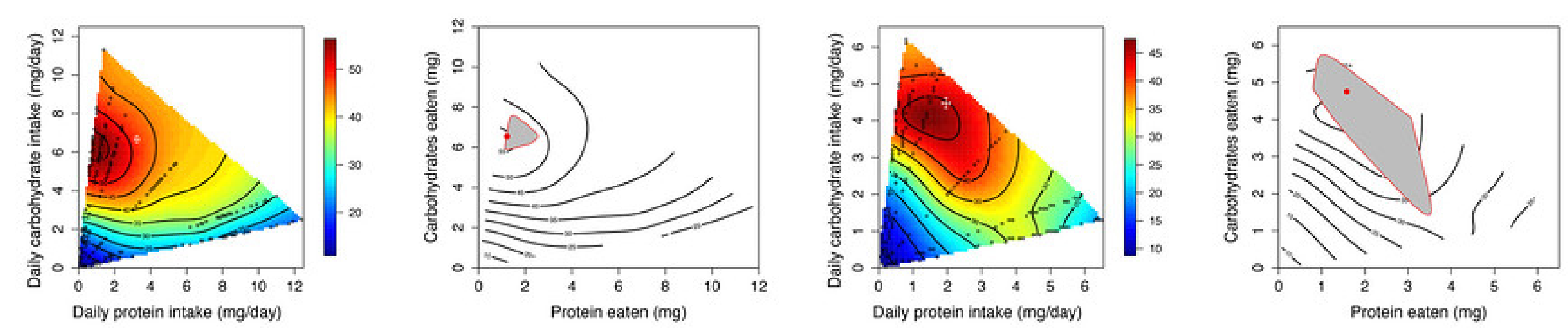


Figure 2. Diet lifetime experiments in *Gryllodes sigillatus* (decorated cricket). A thin plate spline is fit to the lifetime response in females (left) and males (3rd figure) as diets changed proportions and amounts of P and C. Second and fourth plots: corresponding CR's on the location of the optimal diets. The CR of the maximums overlap, providing evidence the optimal diets for female and males are equal. Note the high C to P ratio inside the CR's, away from the origin, so *caloric restriction does not increase lifetime*. Experiments conducted by Prof. John Hunt's group, Biosciences dept., U. of Exeter, UK).

Some selected journal papers from this work

1. Hunt, J., Rapkin, J., and Del Castillo, E., "Evolution of dietary sex differences in Gryllodes sigillatus", to be submitted to **Evolution** (2016).
2. Del Castillo, E., Hunt, J., and Rapkin, J., "OptimCR: an R package for finding confidence regions of response surface optima", to be submitted to **J. of Statistical Software** (2016).
3. Del Castillo, E., Colosimo, B., and Tajbakhsh, S. "Geodesic Gaussian Processes for the Reconstruction of a 3D Free-Form Surface", **Technometrics** (2015).
4. Del Castillo, E., and Colosimo, B.M., "Statistical Shape Analysis of Experiments for Manufacturing Processes", **Technometrics**, (2011).
5. B. Bettonvil, E. del Castillo, and J.P.C. Kleijnen. "Statistical testing of optimality conditions in multiresponse simulation-based optimization," **European J. of Operational Research**, (2009)
6. Cahya, S., Del Castillo, E., and Peterson, J.J., "Computation of Confidence Regions for Optimal Factor Levels in Constrained Response Surface Problems," **J. of Computational and Graphical Statistics**, (2004).
7. Peterson, J., Cahya, S., and Del Castillo, E., "A General Approach to Confidence Regions for Optimal Factor Levels of Response Surfaces", **Biometrics**, (2002).

Acknowledgements

1. National Science Foundation grants DMI 9623669, DMI 9988563, CMMI 0085041, CMMI 0825786 and CMMI 1537987.
2. Ministero dell'Istruzione, dell'Università e della Ricerca (Italy) grant FIRB RBIP069S2T 005.
3. Prof. Bianca Colosimo, Dept. of Production Engineering, Politecnico di Milano, Italy.
4. Dr. John Peterson, Director of Statistics, GlaxoSmithKline.
5. Prof. John Hunt, Evolutionary Genetics Dept., U. of Exeter, UK.