Robust Functional Profile Identification for DSC Thermograms

Amy M. Kwon
SRC, Department of Statistics
Seoul National University
Kwanak-ro 1, Seoul, Korea
amykwon@snu.ac.kr

Dianxu Ren
Department of Biostatistics
University of Pittsburgh
Pittsburgh, USA 15260
dir8@pitt.edu

Ming Ouyang
Department of Computer Science
University of Massachusetts Boston
Boston, USA 02125
ming@cs.umb.edu

Nichola C. Garbett
Department of Internal Medicine
University of Louisville
Louisville, USA 40292
nichola.garbett@louisville.edu

ABSTRACT

Differential scanning calorimetry is an emerging technique with an attempt to characterize a subject's disease status according to heat capacity profiles, which are called thermograms. However, thermograms exhibit large shape variations, and the sample size is typically small. Therefore, it is important to extract robust characterization of thermograms representing the clinical status for further study. The current study identifies the representative heat capacity profiles from functional principle components which are derived from the bootstrap distribution of the deepest heat capacity function according to the functional data depth, instead of the original thermogram data set. 71 thermograms are obtained from two groups (healthy, cervical carcinoma), and functional PCA are conducted with the original thermogram data set and the bootstrap data set of the deepest heat capacity functions. Examining the first three PCs of the two groups between the two data sets, the bootstrap data set shows more distinctive difference in modes of variation between the two groups in comparison with the original thermogram data set, and the representative heat profiles are reconstructed with the PCs which are derived from the bootstrap sample sets. 90% confidence intervals of the representative heat profiles can be directly obtained from the same bootstrap set.

Categories and Subject Descriptors

[**Special Track II**]: Bioinformatics

General Terms

Theory

Keywords

Functional Data Depth, Functional PCA, Bootstrap, Thermograms

1. INTRODUCTION

Thermograms are thermal transition profiles which are generated by measuring the change in heat capacity as a function of temperature in human blood plasma, serum or other biofluid samples using differential scanning calorimetry (DSC) as an analytical method for proteomics [1, 12]. Thermograms have unique profiles reflecting the thermodynamic information of certain molecular constituents in the biological fluid based on structural transitions occurring in response to a temperature change [8, 10]. Thermodynamic information can be represented as the statistical weighting of the Gibbs free energy of all accessible molecular states of a protein in terms of a partition function [10]. Some studies have attempted to characterize the differences between subjects with clinically diagnosed diseases from healthy subjects in terms of the thermal transitions of major proteins by point–wise means or medians [1, 11, 16, 19]. However, there are large shape variations in thermal transition profiles although subjects' clinical states are the same because DSC is sensitive to protein composition and interaction. In addition, the sample size is typically small. Thus, it is critical to extract robust representative heat profiles for sub–populations such as a group of subjects having diagnosed diseases to facilitate further studies. The current study uses the functional data depth instead of the point–wise means or medians. The functional data depth provides a center–outward ordering of a sample of functional curves such as thermograms. The most centered curve is the deepest curve. The current study generates the asymptotic distribution of the deepest heat transition profile using the bootstrap method, and conducts functional principal component analysis (FPCA) with the asymptotic

.

BICT 2015, December 03-05, New York City, United States

DOI 10.4108/eai.3-12-2015.2262549

distribution of the deepest heat transition profile instead of the original thermogram data set. The representative thermal transition profile is reconstructed with a finite number of principal components (PCs). 71 thermograms were obtained from female subjects, and 32 subjects among them were diagnosed with cervical carcinoma. The purpose of the current study is to characterize this thermogram data set for the two groups (healthy controls, cervical carcinoma) based on the bootstrapped functional data depth.

2. METHODS

2.1 Smoothing

A thermogram measures the excess heat capacity (C_p^{ex}) as a function of temperature. Suppose that $\{y(t) : t = 1, \cdots, T\}$ denote a sequence of C_p^{ex}, and t can be defined on a compact interval of $t \in [0, 1]$. The i_{th} thermogram, $y_i(t)$, can be expressed as a function as Eq (1) where $i = 1, \cdots, N$ and $\epsilon_i(t) \sim (0, \sigma_i^2)$.

$$y_i(t) = \mu_i(t) + \epsilon_i(t) \quad (1)$$

The statistical weighting for thermodynamic information is generally made by Gaussian distributions. In that case, Gaussian kernel [20, 24] may be a natural candidate for the estimate of $\hat{\mu}_i(t)$ as $\hat{\mu}_i(t) = \sum_{j=1}^{T} \phi_i(t) \cdot y_j$ where $\phi_i(t) = \frac{K_h(t_j - t)}{\sum_{j=1}^{T} K_h(t_j - t)}$ and h is a tuning parameter or a bandwidth. h can be determined by cross–validation [13].

2.2 The Deepest Function

The Functional Data Depth.

The functional data depth is a real valued functional on a space of functions $[0, 1]$, which indicates the centrality of a function in a given finite cloud of functions [2]. Functional data depths have been proposed in diverse expressions, but there are typically two types [2]. One is the univariate type, and the other is the random projection type. The univariate type calculates the functional data depth by integrating the values of a function over the whole interval such as Fraiman & Muniz (FM) [9] and Graphic Band Depth (GBD) [18]. The random projection type calculates the functional data depth by employing random projection of a univariate depth of a function value at a randomly chosen index value t such as Random Projection (RP) and Random Projection using Derivatives (RPD) [5]. In addition, there are h–mode depth [5], and Kernelized functional spatial depth (KFSD) [23]. One important property of the data depths is maximality at center. As a result, if the supremum of the depths in terms of a sequence of random variables is determined at m_n, m_n almost surely converges to the true center m as the sample size goes to infinity [17, 26]. In the case of univariate variables, m is the median. However, when the sample size is small, this asymptotic property may not hold [2].

Bootstrap.

The bootstrap method has been widely used to approximate the asymptotic distribution of an estimator of interest such as a variance estimator. In particular, distribution approximation of the bootstrap method is valid for quantiles as well as t–statisitcs [3]. In addition, the bootstrap method can provide more accurate inferences when the data are not well behaved or when the sample size is small without reference to external assumptions [7]. The current study draws K sets of bootstrap samples from the original thermogram data set, and generates approximated distribution of the deepest excess heat capacity function. K is 50 in the current study. The deepest excess heat capacity function is determined according to the functional data depths, and four different functional depths (FM, GBD, RP and RPD) were considered to select the deepest excess heat capacity function. A set of the bootstrap samples are generated for $k = 1, \cdots, K$ as follows:

1. Randomly draw $y^{(k)}$ of size N from $\{y_1(t), \cdots, y_N(t)\}$ with replacement.
2. Compute the functional data depth with $y^{(k)}$ according to the four functional data depths: FM, GBD, RP and RPD.
3. Select the deepest excess heat capacity functions according to the four functional data depths where the deepest excess heat capacity functions are denoted as $\{\hat{\theta}_{FM}^{(k)}(t), \hat{\theta}_{GBD}^{(k)}(t), \hat{\theta}_{RP}^{(k)}(t), \hat{\theta}_{RPD}^{(k)}(t)\}$, respectively.

Choice of the functional data depth.

The approximated distributions of the deepest excess heat capacity function can be different by the expressions of the functional data depths. 90% confidence intervals (CIs) are computed with the four approximated distributions of the deepest excess heat capacity function. The original thermogram data set is smoothed by Gaussian kernel at each measurement, so the standard intervals are generated as follows where l ={FM, GBD, RP, RPD}. Other types of bootstrap CIs can be referenced to [6].

1. The bootstrap estimates are computed as $\hat{m}_l(t) = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)}(t)$.
2. Corresponding standard deviations of the bootstrap means are computed as $\hat{se}(m_l(t)) = \sqrt{\frac{1}{(K-1)} \sum_{k=1}^{K} (\hat{\theta}_l^{(k)}(t) - \hat{m}_l(t))^2}$.
3. The standard CIs are computed as $\hat{m}_l(t) \pm z_{1-\frac{\alpha}{2}} \cdot \hat{se}(m_l(t))$ where z denotes standard normal distribution and α is a given false positive value 0.1.

The current study computes the lengths of CIs for the four functional data depths (FM, GBD, RP, RPD), and compares the minimum, the average and the maximum lengths of CIs. The functional data depth is chosen which has the minimum length in the maximum lengths of CIs, and the corresponding distribution of the deepest excess heat capacity function is selected.

2.3 Functional Principal Components

Functional principal component analysis (FPCA) is a key technique to characterize functional curves by exploring features [22]. The frame work of functional PCA is analogous to multivariate PCA. In the functional context, principal components (PCs) are represented by a weight function defined over the given range of t where the weight function is defined to maximize the variance of the principal scores, and PCs are orthogonal among each other. Eq (2) is the frame

equation, and the details to obtain functional PCA can be obtained from [4, 14, 21, 22, 25].

$$\mu_i(t) = \mu(t) + \sum_{p=1}^{\infty} \sqrt{\lambda_p} \cdot \psi_p(t) \cdot \xi_{ip} \quad (2)$$

where $\mu(t)$ is the mean function, $\lambda_k \geq 0$ are eigenvalues of (a finite rank of) $cov(\mu_i(t), \mu_i(s)) \geq 0$, $\psi_p(t)$ are the corresponding orthonormal eigenfunctions and $\{\xi_{ik} : k \geq 1\} \sim (0, 1)$ for each i.

The functional PCA is useful to describe the modes of variation of curves, but PCs can be affected when modulations of some functional curves influence the variance. The variance matrix may be made after discarding some outlying functional curves, but it can cost efficiency. The bootstrap estimate of the deepest functional data depth has been applied as representative curves, and showed better performance in classification than traditional functional data depths [15]. The approximated distribution which is obtained from the bootstrap method replaces the original thermogram data set in the current study. Namely, $\mu_i(t)$ is replaced by $\mu_l^{(k)}(t)$ in Eq (2) to conduct FPCA where l is the selected functional data depth.

2.4 Functional Profile Characterization

The individual thermogram can be predicted by reconstructing $\hat{\mu}_l^{(k)}(t)$ with a finite number of λ_p and $\psi_p(t)$ in Eq (2). The finite number of PCs can be determined according to log likelihood using cross-validation [21]. The reconstructed thermogram set is the predicted bootstrap distribution of the deepest excess heat capacity functions. Consequently, the representative heat capacity functions can be computed by averaging the predicted thermograms, and CIs can be directly computed as described in the last paragraph of Section 2.2.

3. RESULTS

The thermograms were generated as a function of temperature from blood plasma samples of 71 female subjects. About 45% of the subjects were diagnosed with cervical carcinoma, and the thermogram data set was previously used [11]. The observed excess heat capacity was smoothed using Gaussian kernels, and the tuning parameter h was 0.4. All thermograms were labeled as either 'healthy' or 'carcinoma'. 50 sets of bootstrap samples were generated for each group. The deepest excess heat capacity functions were determined per each set of bootstrap samples according to FM, GBD, RP and RPD. (Modified band depths (MBD) were computed for GBD [18]). The bootstrap estimates of the deepest excess heat capacity function with 90% CIs were computed, and their lengths of CIs were compared across the different functional data depths in a group. The results were summarized in Table 1. RP and RPD, which belongs to the random projection type, showed similar lengths of CIs while FM showed smaller lengths of CIs than GBD in the univariate type of functional data depths. The carcinoma group showed bigger lengths of CIs in comparison with the healthy group over all. Consequently, RP was selected which showed the smallest length in the maximum CI lengths based on the results of the carcinoma group. In addition, the point–wise medians, the deepest curve by the functional data depth and the deepest curve by the bootstrap estimate of the deepest curve by the functional data depth were illustrated in Fig 1. The point–wise median curve was less smooth in the healthy group. Also, the deepest curve, which was determined by the typical FM data depth was deviated from typical shapes in the carcinoma group. The bootstrapped functional data depth estimates were smoothly represented in both groups. B–spline based FPCA was conducted with the approximated distribution of the deepest excess heat capacity function which was selected by RP. FPCA was also conducted with the original thermogram data set to compare the modes of variations to those obtained from the approximated distribution of the deepest excess heat capacity function according to RP. Log–likelihood values were compared when the numbers of PCs were from two to five, and the number of PCs was determined as three [21]. The first three PCs were compared between the two data sets. The plots in the upper panels were obtained from the original thermogram data set, those in the lower panel in Fig 2 were obtained from the approximated distribution of the deepest excess heat capacity functions. In Fig 2, the first two PCs accounted for about 98% and 94% of the total variance in the healthy group and the carcinoma group, respectively. The modes of variation by the first two PCs were almost undistinguishable between the healthy group and the carcinoma group in Fig 2. On the other hand, the first two PCs, which accounted for about 94% and 89% of the total variance, were more distinctive in Fig 2. Examining the first PCs, the first higher mode occurred at around $50^{\circ}C$ in the carcinoma group while it occurred at round $60^{\circ}C$ in the healthy group. The second PC began to decrease at around $50^{\circ}C$ in the healthy group while it started to decrease at around $60^{\circ}C$ in the carcinoma group. The approximated distribution was reconstructed based on the first three PCs, and the representative curves and 90% CIs were computed for the two groups. The results were illustrated in Fig 3. The carcinoma group exhibited a lower excess heat capacity of the lower temperature transition than the healthy group. However, the difference in thermal transition was indistinguishable between the two groups above $65^{\circ}C$.

4. DISCUSSION

It is often a primary purpose to characterize the representative thermal transition profile of a specific disease using DSC thermograms. Some previous studies summarized the mean or median excess heat capacity at each measurement point to characterize the thermograms. However, the thermograms have large shape variations and the sample size is generally small. Point–wise means are vulnerable in outlying measurements, and point–wise medians often generate less smooth curves as shown in Fig 1. As a robust measure, the α% trimmed mean may be computed by discarding α% of outlying thermograms based on functional data depths [5]. It can cost efficiency, and it may undesirable when the number of thermograms are small. Instead, the current study generated the approximated distribution of the deepest excess heat capacity function by randomly drawing repeated samples with replacement from the original thermogram data set, which consisted of healthy and cervical carcinoma subjects. The representative thermal transition profiles including CIs were identified by reconstructing the approximated distribution of the deepest excess heat capacity functions based on the finite number of PCs for different clinical groups. We found that the modes in variations were more distinctive between the two different clinical groups

when we used the approximated distribution of the deepest excess heat capacity function. In addition, the representative excess heat capacity function smoothly characterized in comparison with the deepest excess heat capacity function which was computed by a typical functional data depth without losing efficiency. The representative excess heat capacity function was derived from the bootstrap estimates, so it was easier to construct corresponding point–wised CIs. However, it is computationally more intensive than the previous methods, and it still requires further studies for generalization with other thermogram data sets. Also, the performance may be influenced by the choice of the functional data depths.

Table 1: Lengths of confidence intervals (CIs) about the highest depth curves

Depths	The healthy group		
	Min	Mean	Max
GBD	0.0018	0.0318	0.2015
FM	0.0013	0.0248	0.1885
RP	0.0023	0.0353	0.1508
RPD	0.0023	0.0324	0.1499
Depths	The carcinoma group		
	Min	Mean	Max
GBD	0.0036	0.0508	0.2700
FM	0.0017	0.0416	0.2323
RP	0.0036	0.0490	0.1862
RPD	0.0020	0.0537	0.2116

The deepest curves were computed from 50 bootstrap sample sets using four different functional data depths for two groups (Healthy, Carcinoma): GBD [18], FM [9], RP [5], RPD [5]. The lengths of CIs were computed by $(U_i - L_i)$ where $L_i \leq \theta_i \leq U_i$ and θ_i was a parameter of interest from bootstrapping for $i = 1, \cdots, T$. Min and Max indicate the minimum and the maximum lengths of CIs, and Mean indicates the average lengths of CIs.

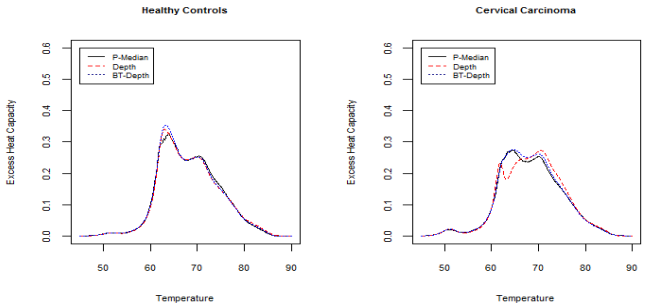

Figure 1: The comparison of the representative curves: 'P–Median' denotes point–wise medians, 'Depth' denotes to the curve of the highest depth based on GBD and 'BT–Depth' denotes to the bootstrap mean of the highest GBD depths

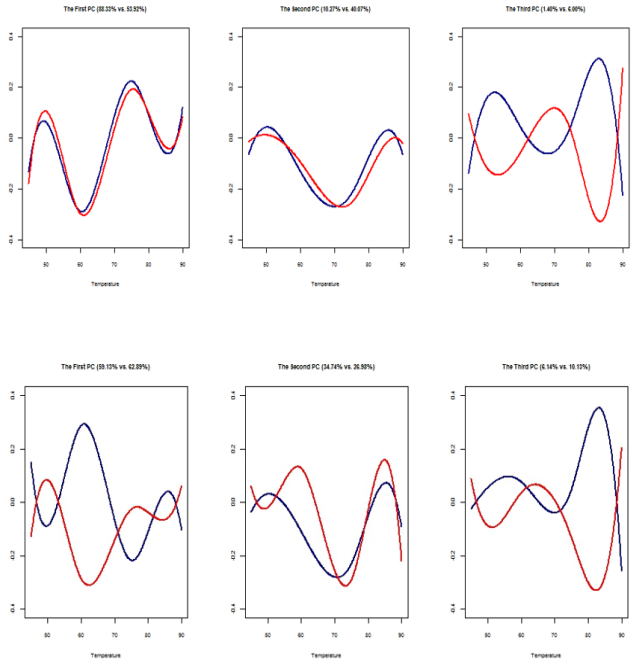

Figure 2: Comparison of the first three principal components between the original data set and the Bootstrap set of RP functional data depth: The plots in the upper panel are the first three PCs from the original data set, and those in the lower panel are the first three PCs

5. ACKNOWLEDGMENTS

Conflict of Interest:NCG is a co-inventor on patent applications describing the DSC thermogram technology. All other authors declare no conflict of interest.

6. REFERENCES

[1] R. Aebersold, L. Anderson, R. Caprioli, and et al. Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res*, 4:1104–1109, 2005.

[2] C. Becker, R. Fried, and S. Kuhnt. *Robustness and Complex Data Structure*. Springer, New York, 2013.

[3] P. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *Ann Stat*, 9:1196–1217, 1981.

[4] G. Boente and R. Fraiman. Kernel–based functional principal components. *Stat Probab Lett*, 48:335–345, 2000.

[5] A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection–based depth notions. *Comp Stat*, 22:481–496, 2007.

[6] T. DiCiccio and B. Efron. Bootstrap confidence intervals. *Stat Sci*, 11:189–228, 1996.

[7] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.

[8] D. Fish, G. Brewood, J. Kim, and et al. Statistical analysis of plasma thermograms measured by differential scanning calorimetry. *Biophys Chem*, 152:184–190, 2010.

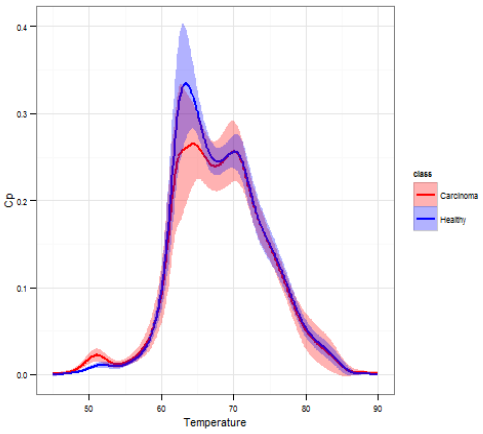

Figure 3: The excess heat capacity function which was reconstructed from functional PCA. The bold lines indicate the representative excess heat capacity functions of the two groups which were computed by averaging the predicted thermograms. The widths are constructed with corresponding 90% CIs.

[9] R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10:419–440, 2001.

[10] E. Freire. Statistical thermodynamic analysis of differential scanning calorimetry data: Structural deconvolution of heat capacity function of proteins. *Methods Enzymol*, 240:502–530, 1994.

[11] N. Garbett, M. Merchant, C. Helm, and et al. Detection of cervical cancer biomarker patterns in blood plasma and urine by differential scanning calorimetry and mass spectrometry. *PLoS*, 9:1–12, 2014.

[12] P. Gill, T. Moghadam, and B. Ranjbar. Differential scanning calorimetry techniques: Applications in biology and nanoscience. *J Biomol Tech*, 21:167–193, 2010.

[13] W. Hardle and J. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Ann Stat*, 13:1465–1481, 1985.

[14] G. James, T. Hastie, and C. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.

[15] A. Kwon, M. Ouyang, and A. Cheng. Resampling based classification using depth for functional curves. *Comm Stat Simul Comp*, 100:1–18, 2014.

[16] L. Liotta and E. Petricoin. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest*, 116:26–30, 2006.

[17] R. Liu. On the notion of data depth based on random simplices. *Ann Stat*, 18:405–414, 1990.

[18] S. Lopez-Pintado and J. Romo. On the concept of depth for functional data. *J Am Stat Assoc*, 104:718–734, 2009.

[19] G. Mor, I. Visinitin, Y. Lai, and et al. Serum protein markers for early detection of ovarian cancer. *Proc Natl Acad Sci*, 102:7677–7682, 2005.

[20] A. Nadaraya. On estimating regression. *Theory Probab Appl*, 9:141–142, 1964.

[21] J. Peng and D. Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J Comput Graph Stat*, 18:995–1015, 2009.

[22] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, 2002.

[23] C. Sguera, P. Galeano, and R. Lillo. Spatial depth–based classification for functional data. *Test*, 23:725–750, 2014.

[24] G. Watson. Smooth regression analysis. *Sankhya*, 26:359–372, 1964.

[25] F. Yao, G. Mueller, and L. Wang. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*, 100:577–590, 2005.

[26] Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann Stat*, 28:461–482, 2000.