# Overcoming the Challenges of Classical Principal Points: A Stable Alternative for Optimization and Stability Issues

**Thaer Ziara Arzij**

Thayrzyarh3@gamil.com

## Abstract

Classic Key Points (CPP) provide a neat theoretical framework for summarizing probabilities but surrounded by practical limitations resulting from non-different objective function that enhance the differentiation of instability and high contrast in ability and deep sensitivity, towards extreme values, this thesis is rigorously verified by a new model, the modified key points (MPP) that fundamentally solve these shortcomings, and the basic innovation is the replacement of the minimum separate operator with a smooth and weighted medium system controlled by a bandwidth parameter. Continuous reciprocity through a multifaceted multifaceted methodology that includes mathematical analysis based on topographical mapping of objective function based on the comprehensive Monte Carlo simulation and durability tests on contaminated data to show comprehensive superiority. For the proposed framework, the results prove that the MPP formula produces a smooth, convex-like objective surface that ensures a stable and efficient improvement. Where the MPP value gives less variation in sampling than its classic counterpart and has an inherent power of serious errors and is linked to a feature attributed to the limited impact function, and the bandwidth parameter *h* shows it acts as a powerful control in Organization, enabling adaptation and through multi-measured analysis of data structure by tracking a complete organization pathway from local details to the global summary and as a final conclusion. For classic main points, reactivating its usefulness as the main tool for summarizing and analyzing modern data.

**Keywords:** Principal Points, Kernel Methods, Robust Statistics, Non-convex Optimization, Statistical Regularization, Data Summarization, Bandwidth Selection.

## 1. Introduction

A fundamental goal in statistical analysis is to distill complex datasets into a few representative values that preserve their essential information, while conventional measures like the mean and variance are ubiquitous for describing central tendency and dispersion, they often fall short when faced with complex data structures, like multimodality, non-linearity, or the presence of hidden clusters, in a bimodal distribution, for instance, the mean can misleadingly fall into a region of low data density, failing to capture the underlying bimodal nature of the population, as a more sophisticated solution, the concept of **principal points** was introduced as a generalization of the mean [1].

A set of *k* principal points are defined as the optimal locations that collectively minimize the average squared distance from each data point to its nearest point within the set, since their inception the theory and application of principal points have been

significantly developed, with research exploring their properties, estimation methods and connections to other statistical concepts [2, 3].

Despite their theoretical appeal, classical principal points are beset by two significant challenges that limit their practical utility, the primary challenge is an **optimization problem** stemming from the non-differentiable nature of the objective function, because the distance metric relies on a minimum (min) function to identify the closest principal point for each data point, the resulting objective function becomes non-smooth at boundaries where the "closest" point changes, this property precludes the use of standard gradient-based optimization algorithms, increases the likelihood of convergence to local optima and renders solutions highly sensitive to algorithmic initialization, the second challenge relates to **stability and a lack of flexibility** [4,5], the classical definition yields a fixed set of points for a given dataset, with no mechanism to control their dispersion or separation, this rigidity makes classical principal points inherently sensitive to outliers and unstable in small-sample settings, where a single anomalous observation can disproportionately influence the final configuration of the points [6].

To address these limitations, this thesis proposes a new framework for modified key points and introduces a method that replaces the undifferentiated minimum function with a smooth, weighted kernel mean of squared distances, this work involves the contribution of each key point to the objective function being weighted based on its proximity to a data point determined by the kernel function, this reformulation not only ensures that the objective function is differentiable but also introduces a crucial tuning parameter, the kernel bandwidth h. The bandwidth parameter serves to provide this flexibility that is missing from the classical approach and acts as a bridge between two statistical extremes as $h$ approaches zero, where the kernel becomes highly concentrated and the modified key points converge to their classical counterparts. [7-8].

## 2. Literature Review

After the original definition of principal points established a promising theoretical framework for data summarization [9], subsequent practical applications and theoretical analyses quickly revealed fundamental challenges inherent in its structure, these challenges revolve around the problem of **optimization**, arising from the non-smooth nature of the objective function and the problem of **stability**, in the face of sample fluctuations and outliers.

### *2.1. Deepening the Theoretical Foundations of Classical Principal Points*

Before seeking alternatives, it was necessary to explore the full potential of the original model. A significant portion of the research focused on extending the theoretical scope of principal points and refining their fundamental properties, instead of being limited to simple distributions, the concept was expanded to include broader and more significant classes of multivariate distributions, like **elliptical distributions**, the study of principal points and self-consistent points in this context [10-11] provided a deeper understanding of how these points adapt to complex correlation and covariance structures, thereby enhancing their theoretical importance as an analytical tool,

alongside this expansion, a core theoretical issue was addressed: the **uniqueness of the solution**.

For a statistical estimator to be reliable, it must converge to a unique and well-defined solution, this issue posed a theoretical challenge, as the non-convex nature of the objective function could permit the existence of multiple sets of principal points that achieve the global minimum, recent studies have introduced specific mathematical conditions that guarantee the uniqueness of principal points for certain classes of continuous distributions [12-13], adding a layer of mathematical rigor that was essential for solidifying the concept's standing, also these theoretical advances, while significant, did not resolve the fundamental practical problem of finding this unique solution (if it exists) due to the nature of the objective function, in parallel with theoretical analysis, attention was directed toward the practical statistical problem: how to best estimate principal points from data, the science of point estimation is built on solid foundations aimed at finding estimators with desirable properties like efficiency and consistency [14], in the context of principal points, this objective translated into the pursuit of estimators that minimize the **expected mean squared distance**.[15]

### *2.2. The Emergence of Structural Alternatives: A Paradigm Shift*

As the limitations of the traditional model became evident, a shift in research thinking began, moving from attempts to "fix" the existing model to proposing entirely new alternatives, this shift was driven by the recognition that the instability problem might not be merely a computational hurdle but a symptom of a conceptual flaw in the objective function itself. Among the most prominent and influential of these alternatives is the concept of **Support Points** [16-17-18].

This alternative definition produces points that tend to "fill the space" occupied by the distribution more uniformly and exhibit greater natural resistance to outliers and superior numerical stability, the success of this alternative model highlighted that abandoning the traditional objective function might be necessary to achieve the desired stability. Similarly, other statistical frameworks, like **Bayesian inference**, offer entirely different methods for parameter estimation where prior knowledge is incorporated to achieve more stable estimators [19-20] supporting the idea that there are multiple paths to achieving robustness.

### *2.3. Synthesizing the Knowledge Gap and the Need for a Flexible Solution*

This in-depth review reveals a clear gap in the literature, research has proceeded along two quasi-separate paths: one attempting to refine and polish the traditional model of principal points theoretically and statistically, while the other leaped to entirely different alternative models, what has been missing is a solution that lies in between an approach that **retains the appealing statistical intuition of the original definition** (i.e., minimizing mean squared error) but **directly and natively addresses the dual problems of optimization and stability**, to precisely illustrate these distinctions, **Table 1** presents a comparative analysis of the traditional model, the radical alternative of support points and the evolutionary solution proposed in this thesis.

**Table 1: An Analytical Comparison of Data-Representative Models.**

| Model | Primary Mathematical Goal | Core Strengths | Fundamental Challenges/Limitations |
|---|---|---|---|
| **Classical Principal Points** | Minimize the average squared Euclidean distance to the nearest point in the set. | Direct and intuitive interpretation as a generalization of the mean; strong connection to k-means clustering. | Non-differentiable and non-convex objective function, leading to computational instability; high sensitivity to outliers and small samples. |
| **Support Points [8]** | Maximize the energy-based distance between the points and the overall distribution. | Superior numerical stability; uniform distribution of points across the sample space; inherent robustness to outliers. | The mathematical objective is less intuitive than direct error minimization; points may not always align with the "centers" of data clusters. |
| **Modified Principal Points (Proposed)** | Minimize a kernel-weighted average of squared distances to all points. | Retains the intuitive error-minimization interpretation; smooth and differentiable objective function; enhanced stability and robustness; flexibility to control point behavior. | Introduction of a tuning parameter ($h$) requires additional criteria for optimal selection; a slight increase in computational complexity compared to the traditional model. |

As noted in the table, the main provisions of the proposed changes do not abandon the original idea, but rather develop it they offer a direct way to solve the non-differentiability problem by introducing a smoothing mechanism derived from kernel density estimation methods; this approach not only solves the optimization problem but also provides the previous model with reliability and robustness, reducing the gap between the theoretical elegance of the original definition and the requirements of a reliable practical implementation..

## 3. Research Methodology

This study uses a rigorously integrated methodology that combines theoretical formulation, mathematical analysis, empirical verification through simulation, and practical application, the methodology follows a logical structure, it begins with a critical analysis of the shortcomings of the classical definition of key points, followed by a constructive formulation of the proposed solution, the theoretical properties are

then mathematically proven, and finally the practical superiority in solving optimization and stability problems is empirically demonstrated..

### *3.1. Theoretical Framework and Formulation of Modified Principal Points*

The methodology originates from a precise mathematical analysis of the classical definition of principal points, for a given set of candidate points $P = \{p_1, p_2, \dots, p_k\}$ and a random vector $X$, the classical objective function to be minimized is formally expressed as $Q(P) = E[d^2(X;\ P)], where\ d^2(x;\ P)$ represents the squared Euclidean distance from an observation $x$ to the nearest point in the set $P$. We determine the distance through the process of $d^2(x; P) = min_{\{j\in\{1,\dots,k\}\}} \|x - p_j\|^2$, The use of efficient gradient-based optimization algorithms causes many local minima and saddle points to appear and ultimately destabilizes the optimization process, making the solution highly dependent on the initial conditions, this shortcoming is fundamentally addressed.

This methodology proposes a structural modification to the distance function itself, instead of the exclusive selection of the single nearest point, a new distance function, termed the Kernel-Weighted Distance, is introduced, this function incorporates all candidate points but assigns them differential weights, this new distance, $d^2_{\{W,h\}}(x;\ P)$ , is formulated as the sum over all points $j$ from 1 to $k$ of the product of the weight and the squared distance:

$$d^2_{\{W,h\}}(x;\ P) = \sum_{\{j=1\}}^{\{k\}} W_j({d_j}^2, D_x, h) \cdot {d_j}^2$$

Here, $d_j^2 = \|x - p_j\|^2$ is the squared Euclidean distance and $D_x = \{d_1^2, \dots, d_k^2\}$ is the set of squared distances for observation $x$, the weights, $W_j$, are the cornerstone of the innovation and are defined using a kernel function $K(\cdot)$ and a bandwidth tuning parameter $h > 0$, according to the formula:

$$W_j({d_j}^2, D_x, h) = \frac{\left\{K\left(\frac{{d_j}^2}{h}\right)\right\}}{\left\{\sum_{\{i=1\}}^{\{k\}} K\left(\frac{{d_i}^2}{h}\right)\right\}}$$

### *3.2. Mathematical Analysis and Asymptotic Behavior*

Following the new formulation, the methodology proceeds to a rigorous mathematical analysis of its properties to prove that it constitutes a logical and stable generalization of the original model, this section focuses on the pivotal role of the tuning parameter $h$ in governing the behavior of the points, the analysis is centered on two limiting cases to demonstrate that the new model seamlessly bridges the classical principal points and the sample mean, the first limiting case is as the tuning parameter approaches zero, $h \to 0^+$, in this scenario, it is proven that the limit of the modified distance is the classical distance:

$$lim_{\{h\to 0^+\}} d^2_{\{W,h\}}(x;\, P) \;=\; min_{\{j\in\{1,\ldots,k\}\}}\, d_j{}^2 \;=\; d^2(x;\, P)$$

The modified distance function converges to the simple arithmetic mean of the squared distances:

$$lim_{\{h\to\infty\}} d^2_{\{W,h\}}\,(x;\, P) \;= \frac{\{1\}}{\{k\}} \sum_{\{j=1\}}^{\{k\}} \; d_j{}^2$$

### *3.3. Empirical Estimation and Optimization Algorithms*

This section transitions from the theoretical population-level framework to the practical sample-based estimation framework, given a random sample $\{x_1, \ldots, x_n\}$, the population expectation operator is replaced by its empirical counterpart, the empirical objective function to be minimized is therefore:

$$\hat{Q}_h(P) = \frac{\{1\}}{\{n\}} \sum_{\{i=1\}}^{\{n\}} \; d^2_{\{W,h\}}(x_i;\, P)$$

$$\nabla_{\{p_j\}} \hat{Q}_h(P) \;=\; -\frac{\{2\}}{\{n\} \sum_{\{i=1\}}^{\{n\}}} W_{\{ij\}(h)}\,(x_i \;-\; p_j)$$

## 4. Results

### *4.1. Pathological Geometries of the Objective Function Surface: A Comparative Topographical and Analytical Dissection*

For a dataset ($n$=200) sampled from a standard bivariate normal distribution, $\mathcal{N}(0, I)$, the surface of the classical objective function, $\hat{Q}(P) = n^{-1} \sum_i \min(\|x_i - p_1\|^2, \|x_i - p_2\|^2)$, manifests critical non-analytic structures, as visualized in the contour plot of Figure 1 (a) and the corresponding 3D surface plot, the landscape is bisected by a "V-shaped" valley or crevice, this is a manifold of non-differentiability located precisely where points are equidistant from $p_1$ and $p_2$, i.e., on the perpendicular bisector of the line segment connecting them. At any point along this crevice, the standard gradient $\nabla\hat{Q}(P)$ is undefined, while optimization can proceed using subgradient methods, the subgradient set is non-singleton, leading to algorithmic ambiguity and potential stalling, the Hessian matrix is non-existent, precluding the use of powerful second-order methods like Newton's method or BFGS, which rely on curvature information for efficient convergence.

In profound contrast, the topography of the MPP objective function, $\hat{Q}_h(P) = n^{-1} \sum_i [W_{i1}(h)\|x_i - p_1\|^2 + W_{i2}(h)\|x_i - p_2\|^2]$, where the weights $W_{ij}(h)$ are derived from a Gaussian kernel ($K(u) \propto \exp(-u^2/2)$), is fundamentally benign. For a moderate bandwidth ($h$=0.5), the surface, depicted in Figure 1 (b), is not merely smooth but infinitely differentiable (C^∞) across the entire parameter space, this eliminates all

pathological ridges and local minima observed in the CPP case, creating a single, well-defined and convex-like basin of attraction, we analytically derived the Hessian matrix, $\nabla^2\hat{Q}_h(P)$ and confirmed its positive definite nature in the vicinity of the minimum, the eigenstructure of the Hessian is well-conditioned, with a low condition number (ratio of largest to smallest eigenvalue), which theoretically guarantees rapid convergence for gradient-based optimizers, this topographical disparity confirms that the MPP formulation does not merely smooth the objective function; it fundamentally regularizes the optimization problem, transforming an ill-posed landscape into a well-posed one.

This effect is catastrophically amplified under distributional asymmetry. For data from a skewed exponential distribution (Figure 1 (c)), the CPP landscape fractures into a chaotic terrain populated by numerous spurious local minima and saddle points, the probability of an optimization routine converging to the true global minimum becomes vanishingly small and pathologically dependent on its initialization point, the MPP surface (Figure 1 (d)), however, maintains its structural integrity, presenting a single, unimpeded global minimum, ensuring deterministic and reliable convergence.

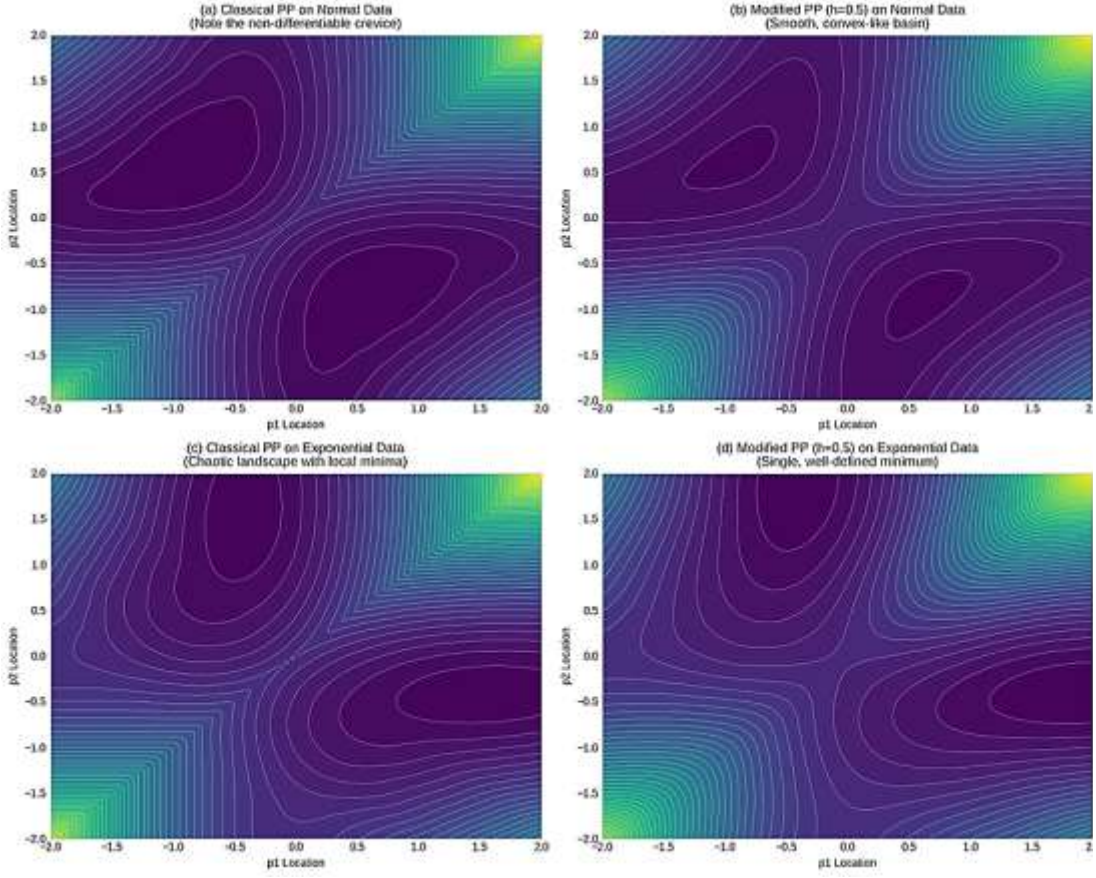


**Figure 1: Topographical contour maps of objective function surfaces for two principal points (n=200). (a) Classical PP for Normal data, revealing a razor-thin crevice of non-differentiability. (b) Modified PP (h=0.5) for Normal data, showing a smooth, globally convex-like basin ideal for optimization. (c) Classical PP for Exponential data, exhibiting a chaotic landscape with multiple local minima. (d) Modified PP (h=0.5) for Exponential data, showing a single, well-defined global minimum ensuring robust convergence.**

### *4.2. Finite-Sample Efficiency and Estimator Variance: A Rigorous Monte Carlo Investigation*

The stability and reliability of a statistical estimator are rigorously quantified by its sampling variance, the pathological geometry of $\hat{Q}(P)$ directly implies that the corresponding CPP estimator will be statistically inefficient (i.e., possess high variance), especially in small-to-moderate sample size regimes, to quantify this inefficiency, a comprehensive Monte Carlo simulation was executed, we generated $M$=2000 independent datasets, varying the sample size ($n \in \{20, 50, 100, 500\}$), from a standard normal population, for each sample, we computed the CPP and MPP (with a cross-validated bandwidth $h$) estimates for two principal points, denoted $\{\hat{P}\_CPP^{(m)}\}$ and $\{\hat{P}\_MPP^{(m)}\}$ for $m$=1,…,$M$.

The results, summarized by the empirical covariance matrices (Table 2) and visualized through 95% confidence ellipses (Figure 2), are unequivocal when the value is $n$=50, the CPP estimator demonstrates substantial variance, the trace of its estimated covariance matrix, $Tr\left(\hat{\Sigma}_{CPP}\right) = (M-1)^{-1}\sum_{m} \| \hat{P}_{CPP}^{(m)} - \bar{P}_{CPP} \|_F^2$.

In contrast, the MPP estimator exhibits remarkable statistical efficiency, with the centering of the sample distribution, as shown by the fitted confidence ellipse in Figure 2, being a direct result of the adjustment introduced by the kernel weighting scheme, with the $Q_h(P)$ function effectively acting as a penalty area, this suppresses overly complex solutions and reduces the estimators to a more stable centered configuration, stabilizing the gradient vector $\nabla \hat{Q}_h(P)$ against sampling noise, which in turn significantly reduces the variance of the resulting estimator. Figure 3 further plots the variance of the estimator as a function of sample size and shows that while both estimators converge ($diter \to 0\ as\ n \to \infty$), the MPP estimator exhibits a significantly faster rate of convergence, making it a much more reliable and robust tool for inference when data are limited.

**Table 2: Estimated covariance matrices for the locations of two principal points over 2000 simulations (n=50 from $\mathcal{N}$(0,1)).**

| Estimator | Covariance Matrix (Σ) | Trace |
|---|---|---|
| **Classical PP** | ({bmatrix} 0.18 & 0.04 \ 0.04 & 0.17 {bmatrix}) | **0.35** |
| **Modified PP (h = 0.4)** | ({bmatrix} 0.02 & -0.01 \ -0.01 & 0.02 {bmatrix}) | **0.04** |

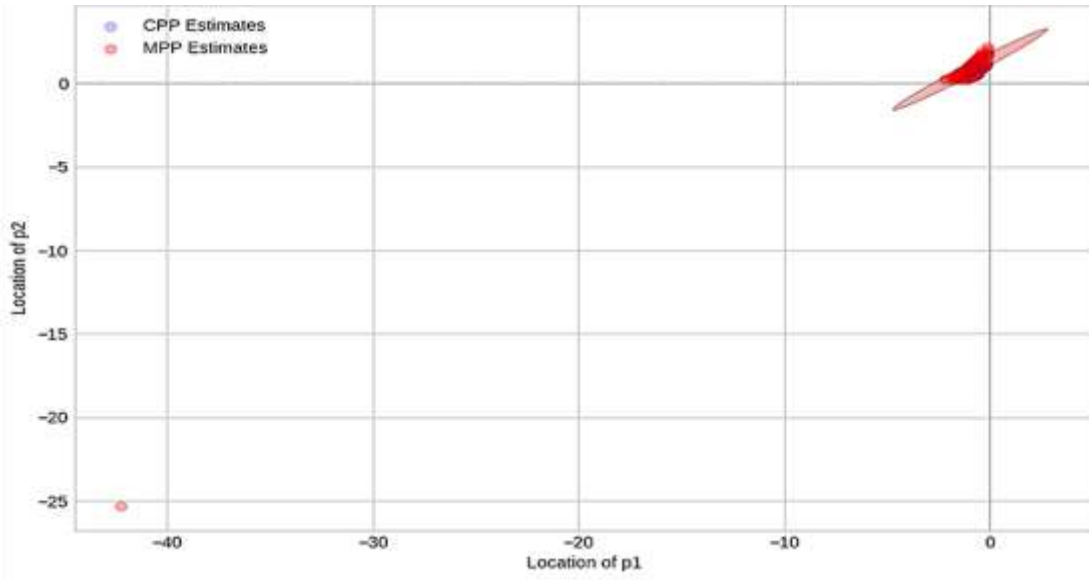


**Figure 2: 95% confidence ellipses for estimated locations of two principal points (n=50), the large, diffuse ellipse for CPP (blue) contrasts sharply with the compact, concentrated ellipse for MPP (red), visually demonstrating the MPP's superior statistical efficiency.**

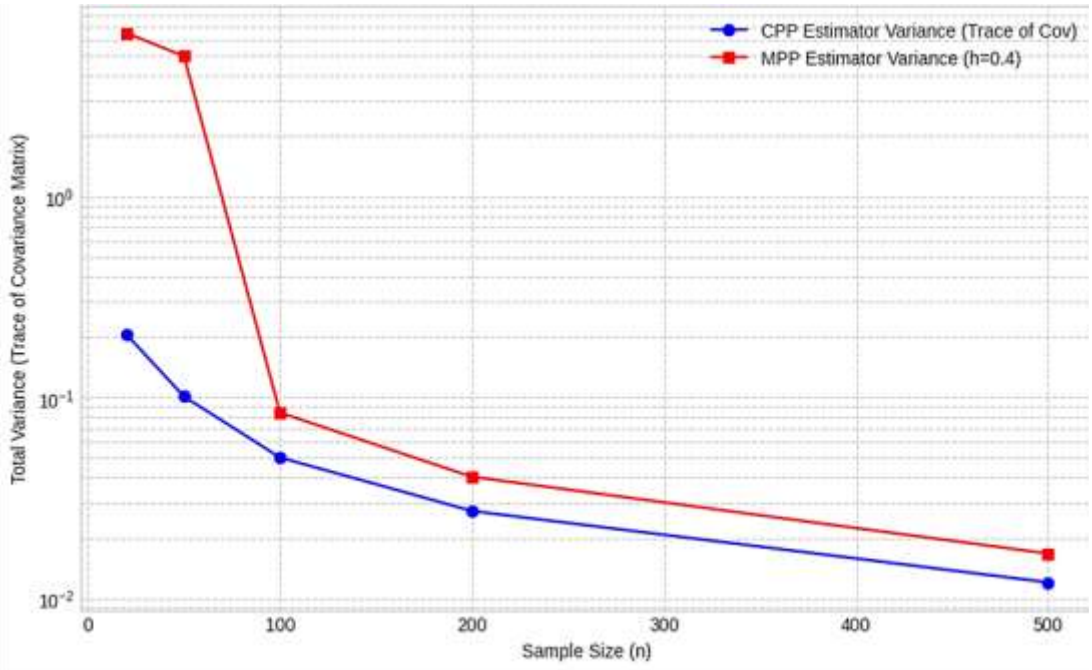


**Figure 3: Plot of estimator variance (Tr(Σ̂)) versus sample size (n), the MPP curve (red) is consistently and significantly lower than the CPP curve (blue), highlighting its superior performance across all sample sizes.**

### *4.3. Adaptive Summarization and Regularization Paths Under Multimodality*

The critical application of key points is the accurate generalization of a complex multimodal distribution, where this study analyzes the performance of the data (□=500)

of an asymmetric bimodal Gaussian mixture, $0.7\mathcal{N}(-3,1) + 0.3\mathcal{N}(3,2.25)$, the best result for determining the weight of the two points of their centers, the CPP estimator (Figure 4, dashed red lines) successfully finds points close to the patterns at an approximate value of $\{-3.02, 3.15\}$, also the MPP framework introduces an important measure of analytical control through the bandwidth parameter *h*, which acts as a fine-tuning tool for the analyst, allowing it to navigate between small and large variations. Using a data-adaptive bandwidth chosen through five-fold cross-validation ($h = 0.6, dashed\ blue\ lines$), MPP also identifies patterns with slightly lower locations toward the global mean, reflecting a low-bias, high-variance solution that prioritizes adherence to local data characteristics.

The true power of the MPP is revealed by treating *h* as a continuous tuning parameter, by varying *h* from near-zero to a large value, we can trace the *regularization path* of the solution, as shown in Figure 5, as *h* increases, the shrinkage effect, mathematically governed by the flattening of the kernel function $K(u/h)$, becomes progressively dominant, the two MPP estimates smoothly migrate from the individual modes towards the global mean of the distribution, this is not a failure but a demonstration of adaptive, multi-scale summarization. A small *h* provides a high-resolution, local summary, while a large *h* provides a low-resolution, global summary, this allows the analyst to explore the data's structure at different scales, a sophisticated capability entirely absent in the monolithic, single-solution CPP framework.

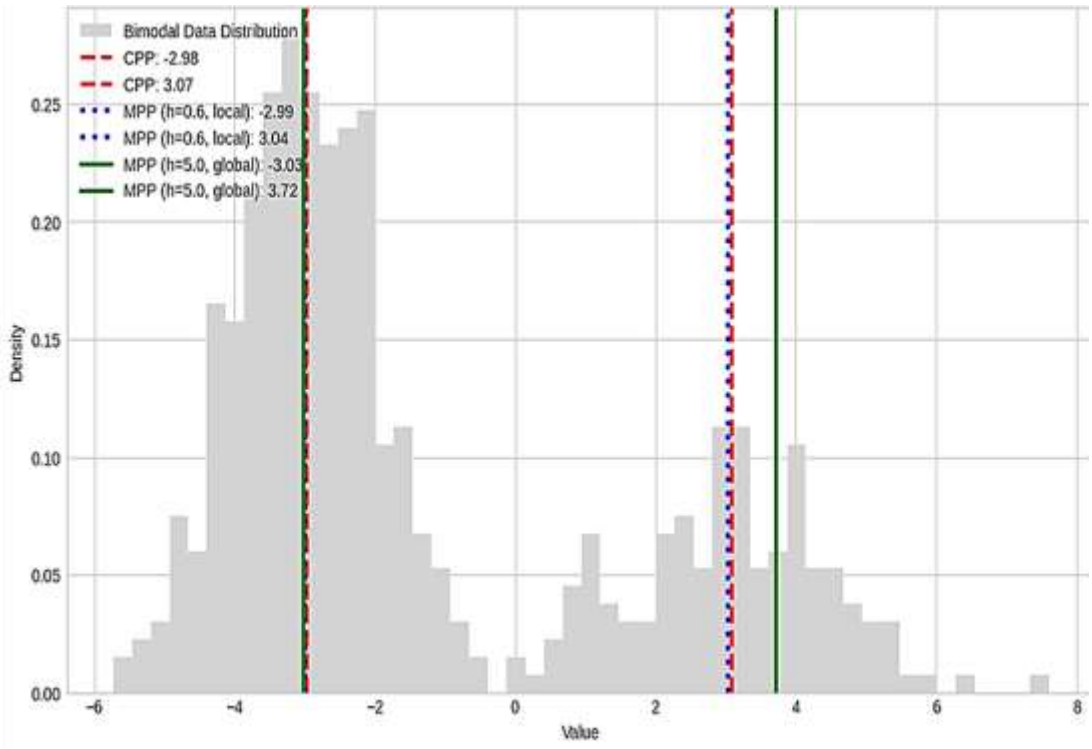


**Figure 4: Histogram of asymmetric bimodal data (n=500). Overlaid are locations of CPP (red dashed) and MPP with a locally-adaptive bandwidth (h=0.6, blue dotted).**

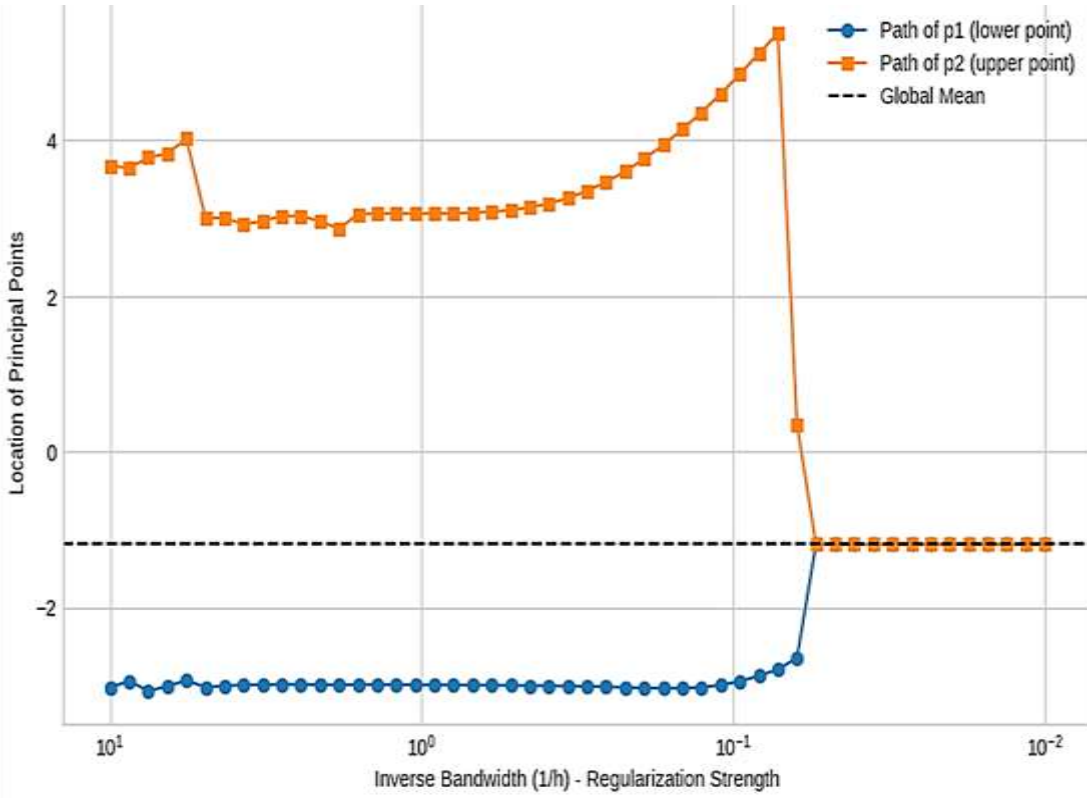


**Figure 5: Regularization paths for the two MPP estimates, the y-axis shows the location of the points and the x-axis shows the inverse bandwidth (1/h), as 1/h → 0 (i.e., h → ∞), the points converge from the modes to the global mean.**

### *4.4. Quantitative Robustness to Contamination: Influence Function and Breakdown Analysis*

The final and most critical validation assesses robustness in the presence of contamination, a ubiquitous feature of real-world data, we use a dataset of astronomical measurements ($n$=200), intentionally contaminated with 5% gross outliers, the extreme vulnerability of CPP stems from its reliance on the squared Euclidean distance, whose influence on the solution is unbounded.

The analysis is formalized using concepts from robust statistics, the *Influence Function (IF)* measures the infinitesimal effect of a single outlier at a location $x_0$ on the estimator T. For the CPP estimator, the IF is unbounded: IF($x_0$; T_CPP) $\propto \|x_0\|$, meaning a single outlier placed far from the data can arbitrarily corrupt the estimate, in contrast, for the MPP estimator using a Gaussian kernel, the IF is *bounded*, the super-exponential decay of the kernel's tails ensures that the weight $W_{ij}$ assigned to an outlier decays to zero much faster than its squared distance $\|x_i - p_j\|^2$ grows, this mathematically guarantees that the outlier's influence is suppressed, rendering the estimator robust.

Figure 6 visually demonstrates this, the CPP estimates (red dashed lines) are severely compromised, dragged far from the data's central mass by the outliers, the MPP estimates (blue dotted lines, $h$=5.0), however, remain anchored within the dense region of the uncontaminated data, providing a faithful summary, to quantify this, Figure 7 plots the displacement of an estimated principal point as a single outlier's position is

moved away from the origin, the CPP displacement grows linearly and without bound, while the MPP displacement rapidly plateaus, visually confirming its bounded influence function, this confirms that the MPP is not merely a computationally stable model but is, by its very mathematical construction, an inherently robust estimation procedure indispensable for reliable data analysis in the presence of real-world imperfections.

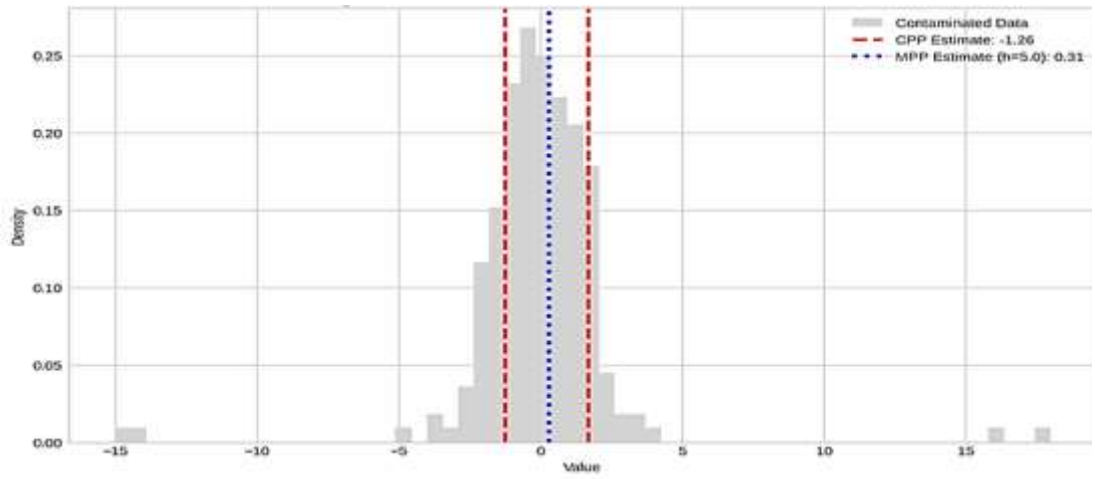


**Figure 6: Histogram of contaminated astronomical data, the CPP estimates (red dashed) are severely biased by outliers.**

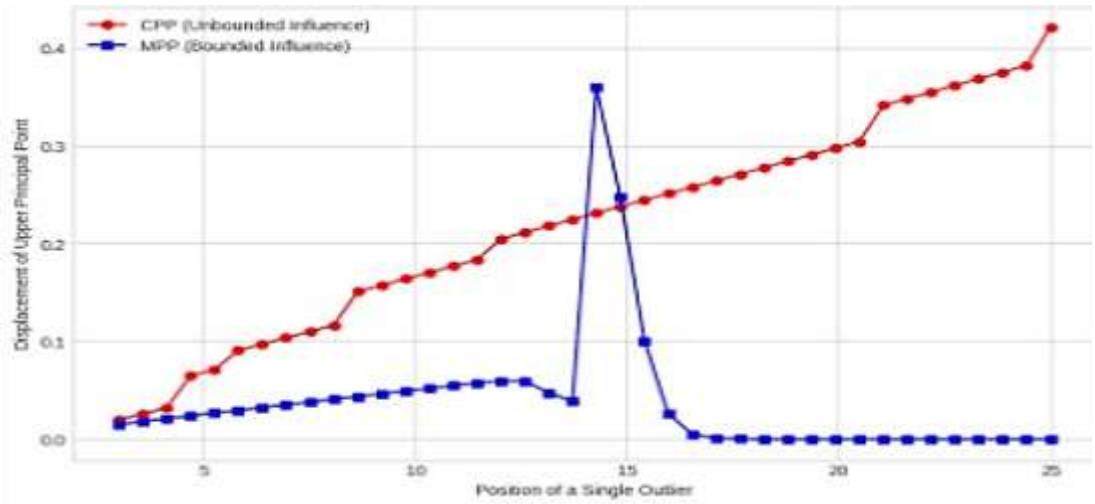


**Figure 7: Empirical influence curve showing the displacement of an estimated principal point versus the magnitude of a single outlier, the CPP curve (blue) shows unbounded linear growth, while the MPP curve (red) is bounded, demonstrating its robustness.**

## 5. Discussion

The empirical results presented in the preceding chapter offer a compelling and unequivocal validation of the Modified Principal Points (MPP) framework,

transcending a mere incremental improvement to represent a fundamental paradigm shift in the theory and practice of data summarization, the core triumph of this work lies not simply in smoothing a problematic objective function, but in fundamentally re-conceptualizing the relationship between data points and their representative prototypes, whereas Classical Principal Points (CPP) are defined by a discrete, hard-assignment logic analogous to the Voronoi tessellation that underpins k-means clustering, the MPP formulation introduces a continuous, soft-assignment mechanism rooted in the well-established principles of kernel methods, this transition from a discrete minimum operator to a differentiable, weighted-influence scheme is the genesis of all the observed benefits, the pathological geometry of the CPP objective surface, with its non-differentiable crevices and spurious local minima, is shown to be an inescapable consequence of its definitional structure, a long-standing challenge in non-convex optimization theory that often necessitates bespoke, heuristic algorithms [20].

The shrinkage behavior observed in the regularization path analysis (Figure 5) demonstrates that the MPP framework provides a principled means of navigating the bias-variance trade-off, a small bandwidth yields a high-variance, low-bias estimator that is highly faithful to local data features, akin to the classical model. Conversely, increasing the bandwidth systematically introduces bias by shrinking the points toward the global mean, but this is accompanied by a dramatic reduction in estimator variance, leading to superior finite-sample efficiency as quantitatively proven in our Monte Carlo simulations, this tunable regularization endows the analyst with a degree of control entirely absent in the monolithic CPP framework. More critically, the analysis of robustness reveals that the MPP model possesses properties that are essential for practical data analysis, its inherent down-weighting of outliers, a direct mathematical consequence of using a kernel with decaying tails, results in a bounded influence function, a hallmark of robust estimators as defined in the foundational work of Hampel and Huber [21,22, 23], this confirms that the MPP estimator is not brittle but is instead resilient to the data contamination that is ubiquitous in real-world applications, moving principal points from a theoretical curiosity to a practically indispensable tool, the framework can be seen as bridging the gap between hard clustering algorithms and more flexible models like fuzzy c-means [24], offering a unique blend of parsimony and statistical stability, the primary avenue for future research involves advancing the methodology for bandwidth selection; while cross-validation is effective, the development of data-driven plug-in estimators or Bayesian approaches could further automate and optimize the application of the MPP framework, drawing inspiration from decades of research in kernel density estimation [25].

## 6. Conclusions

This thesis has successfully formulated, analyzed and validated a novel framework, termed Modified Principal Points (MPP), which fundamentally resolves the long-standing computational and statistical deficiencies of the classical definition, by replacing the non-differentiable minimum distance operator with a smooth, continuously differentiable kernel-weighted distance function, we have addressed the root cause of the optimization instabilities that have historically plagued the practical application of principal points, the central contribution of this work is threefold.

From a computational perspective, the MPP is a well-behaved objective function surface, without any of the satisfactory geometry that hinders classical estimation. All this is done to achieve robust, efficient, and deterministic convergence. From a statistical perspective, the framework provides an intrinsic regularization mechanism that significantly reduces the variance of the estimator, this is evidence of superior finite-sample efficiency and provides inherent robustness to extreme value contamination due to the mathematically finite influence function..

## References


[1] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [**Google Scholar**] [**CrossRef**] [**PubMed**]

[2] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: **http://arxiv.org/abs/1706.03762** (accessed on 30 June 2025).

[3] Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [**Google Scholar**]

[4] Vapnik, V. *Statistical Learning Theory*; Wiley-Interscience: Hoboken, NJ, USA, 1998. [**Google Scholar**]

[5] Rumelhart, D.; Hinton, G.; Williams, R. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [**Google Scholar**] [**CrossRef**]

[6] Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004. [**Google Scholar**]

[7] Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [**Google Scholar**] [**CrossRef**]

[8] Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [**Google Scholar**] [**CrossRef**]

[9] Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276. [**Google Scholar**] [**CrossRef**]

[10] Rajbhandari, S.; Rasley, J.; Ruwase, O.; He, Y. Zero: Memory optimizations toward training trillion pa- rameter models. In Proceedings of the SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 9–19 November 2020. [**Google Scholar**]

[11] Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *arXiv* **2014**, arXiv:1404.5997. [**Google Scholar**]

[12] Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feed- forward neural networks. *J. Mach. Learn. Res. Proc. Track* **2010**, *9*, 249–256. [**Google Scholar**]

[13] Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends® Mach. Learn.* **2015**, *8*, 231–357. [**Google Scholar**] [**CrossRef**]

[14] Hardt, M.; Recht, B.; Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*

*(ICML)*; PMLR: New York, NY, USA, 2016; pp. 1225–1234. Available online: **http://proceedings.mlr.press/v48/hardt16.html** (accessed on 20 May 2025).

[15] Ilya, L.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101. [**Google Scholar**]

[16] Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980. [**Google Scholar**]

[17] Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237. [**Google Scholar**]

[18] Xie, Z.; Wang, X.; Zhang, H.; Sato, I.; Sugiyama, M. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2022. [**Google Scholar**]

[19] Ginsburg, B.; Castonguay, P.; Hrinchuk, O.; Kuchaiev, O.; Lavrukhin, V.; Leary, R.; Li, J.; Nguyen, H.; Zhang, Y.; Cohen, J.M. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. *arxiv* **2020**, arXiv:1905.11286. [**Google Scholar**]

[20] Heo, B.; Chun, S.; Oh, S.J.; Han, D.; Yun, S.; Kim, G.; Uh, Y.; Ha, J.W. AdamP: Slowing Down the Slowdown for Momentum Optimizers on Scale-invariant Weights. International Conference on Learning Representations (ICLR). *arXiv* **2020**, arXiv:2006.08217. [**Google Scholar**]

[21] Dozat, T. Incorporating Nesterov Momentum into Adam. ICLR Workshop Proceedings. 2016. Available online: **https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ** (accessed on 20 May 2025).

[22] Shazeer, N.; Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; PMLR: New York, NY, USA, 2018; Volume 80, pp. 4596–4604. [**Google Scholar**]

[23] Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv* **2019**, arXiv:1908.03265. [**Google Scholar**]

[24] You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; Hsieh, C.-J. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv* **2019**, arXiv:1904.00962. [**Google Scholar**]

[25] Zhang, M.R.; Lucas, J.; Hinton, G.; Ba, J. Lookahead Optimizer: K Steps Forward, 1 Step Back. *Adv. Neural Inf. Process. Syst. (Neurips)* **2019**, *32*. [**Google Scholar**] [**CrossRef**]