The Annals of Statistics
2016, Vol. 44, No. 5, 1888–1895
DOI: 10.1214/16-AOS1493
© Institute of Mathematical Statistics, 2016

PETER HALL'S WORK ON HIGH-DIMENSIONAL DATA AND CLASSIFICATION

BY RICHARD J. SAMWORTH[1]

University of Cambridge

In this article, I summarise Peter Hall's contributions to high-dimensional data, including their geometric representations and variable selection methods based on ranking. I also discuss his work on classification problems, concluding with some personal reflections on my own interactions with him. This article complements [*Ann. Statist.* **44** (2016) 1821–1836; *Ann. Statist.* **44** (2016) 1837–1853; *Ann. Statist.* **44** (2016) 1854–1866 and *Ann. Statist.* **44** (2016) 1867–1887], which focus on other aspects of Peter's research.

1. High-dimensional data. Peter Hall wrote many influential works on high-dimensional data, though notably he largely eschewed the notions of sparsity and penalised likelihood that have become so popular in recent years. Nevertheless, he was interested in variable selection, and wrote several papers that involved ranking variables in some way. Perhaps his most well-known papers in this area, though, concern geometrical representations of high-dimensional data.

1.1. *Geometric representations of high-dimensional data.* Hall and Li (1993) were early pioneers of high-dimensional data analysis in trying to understand the properties of low-dimensional projections of a high-dimensional isotropic random vector X in $\mathbb{R}^p$. As motivation, let $\gamma \in \mathbb{R}^p$ have $\|\gamma\| = 1$ and suppose that

$$\forall b \in \mathbb{R}^p, \exists \alpha_b, \beta_b \in \mathbb{R}, \qquad \mathbb{E}(b^T X | \gamma^T X = t) = \alpha_b t + \beta_b. \tag{1}$$

This condition says that the regression function of $b^T X$ on $\gamma^T X$ is linear. Then, using the isotropy of X,

$$0 = \mathbb{E}(b^T X) = \mathbb{E}\{\mathbb{E}(b^T X | \gamma^T X)\} = \mathbb{E}(\alpha_b \gamma^T X + \beta_b) = \beta_b.$$

Moreover,

$$b^T \gamma = \mathrm{Cov}(b^T X, \gamma^T X) = \mathbb{E}\{\mathbb{E}(b^T X X^T \gamma | \gamma^T X)\} = \alpha_b \gamma^T \mathbb{E}(X X^T)\gamma = \alpha_b,$$

and we conclude that $\mathbb{E}(X|\gamma^T X = t) = t\gamma$, or equivalently,

$$\|\mathbb{E}(X|\gamma^T X = t)\|^2 - t^2 = 0. \tag{2}$$

Received June 2016.

[1] Supported by an EPSRC Early Career Fellowship and a Philip Leverhulme prize.

MSC2010 subject classifications. 01A70, 62-03.
Key words and phrases. Classification, high-dimensional data.

The left-hand side of (2) is always nonnegative, so can be used as a measure of the extent to which the condition (1) holds. Remarkably, under very mild conditions on the distribution of X, Hall and Li (1993) proved that if γ is drawn from the uniform distribution on the unit Euclidean sphere in $\mathbb{R}^p$, then

$$\|\mathbb{E}(X|\gamma, \gamma^T X = t)\|^2 - t^2 \xrightarrow{p} 0$$

as $p \to \infty$. This is equivalent to the statement

$$\sup_{b \in \mathbb{R}^p : \|b\|=1, b^T\gamma=0} |\mathbb{E}(b^T X|\gamma, \gamma^T X = t)| \xrightarrow{p} 0$$

as $p \to \infty$. See also Diaconis and Freedman (1984), who showed that under mild conditions, most low-dimensional projections of high-dimensional data are nearly normal. Of course, when X has a spherically symmetric distribution, (1) holds for every $\gamma \in \mathbb{R}^p$ with $\|\gamma\| = 1$. But the result of this paper shows that even without spherical symmetry, there is a good chance (in the sense of random draws of γ as described above) that (1) holds, at least approximately, when p is large. An important statistical consequence of this is that even if the relationship between a response Y and a high-dimensional predictor is nonlinear, say $Y = g(\gamma^T X, \varepsilon)$ for some unknown link function g and error ε, standard linear regression procedures can often be expected to yield an approximately correct estimate of γ up to a constant of proportionality. The generalisation of this result that replaces $\gamma^T X$ with $\Gamma^T X$, where Γ is a random $p \times k$ matrix with orthonormal columns, also plays an important role in justifying the use of sliced inverse regression for dimension reduction [Li (1991)].

Another seminal paper that articulated many of the key geometrical properties of high-dimensional data is Hall, Marron and Neeman (2005). This paper begins with the simple, yet remarkable, observation that if $Z \sim N_p(0, I)$, then $\|Z\| = p^{1/2} + O_p(1)$ as $p \to \infty$. Thus, data drawn from this distribution tend to lie near the boundary of a large ball. Similarly, the pairwise distances between points are almost a deterministic distance apart, and the observations tend to be almost orthogonal. In fact, the authors go on to explain that, under much weaker assumptions than Gaussianity, the data lie approximately on the vertices of a regular simplex, and that the stochasticity in the data essentially appears as a random rotation of this simplex. As well as clarifying the relationship between Support Vector Machines [e.g., Christianini and Shawe-Taylor (2000)] and Distance Weighted Discrimination classifiers [Marron, Todd and Ahn (2007)] in high dimensions, the paper forced researchers to rewire their intuition about high-dimensional data, and precipitated a flood of subsequent papers on high-dimensional asymptotics.

1.2. *Variable selection and ranking.* The last 15 years or so have seen variable selection emerge as one of the most prominently studied topics in Statistics. Although Peter's instinct was to think nonparametrically [Cheng and Fan (2016), Delaigle (2016), Müller (2016)], he realised that he could contribute to a prominent line of research in the variable selection literature, namely marginal screening [e.g., Fan and Lv (2008), Fan, Samworth and Wu (2009), Li, Zhong and Zhu (2012)], via the deep understanding he developed for rankings. Hall and Miller (2009a) defined variable rankings through their generalised correlation with a response, while Delaigle and Hall (2012) studied variable transformations prior to ranking based on correlation as a method for dealing with heavy-tailed data. For classification, Hall, Titterington and Xue (2009a) proposed a cross-validation based criterion for assessing variable importance, while in the unsupervised setting, Chan and Hall (2010) suggested ranking the importance of variables for clustering based on nonparametric tests of modality.

These works above were underpinned by Peter's realisation that he could explain how perhaps his favourite tool of all, namely the bootstrap [Chen (2016)], could be used to quantify the authority of a ranking [Hall and Miller (2009b)]. In fact, there are some subtle issues here, particularly surrounding the issue of ties. Peter developed an ingenious method for proving that even though the standard n-out-of-n bootstrap does not handle this issue well, the m-out-of-n bootstrap overcomes it in an elegant way.

2. Classification problems. I believe that Peter may have become interested in classification problems in the early 2000s at least partly through ideas of bootstrap aggregating, or bagging [Breiman (1996)]. Indeed, in Friedman and Hall (2007), a preprint of which was already available in early 2000, Peter had attempted to understand the effect of bagging in M-estimation problems. This is a typical example of Peter's extraordinary ability to explain empirically observed effects through asymptotic expansions. One of the other interesting contributions of this work is that subsampling (i.e., sampling without replacement) half of the observations closely mimics ordinary n-out-of-n bootstrap sampling, a very useful fact that has been observed and exploited in several other contexts, including stability selection for choosing variables in high-dimensional inference [Meinshausen and Bühlmann (2010), Shah and Samworth (2013)] and stochastic search methods for semiparametric regression [Dümbgen, Samworth and Schuhmacher (2013)].

Classification problems are ideally suited to bagging, because the discrete nature of the response variable means that small changes to the training data can often yield different outputs from a classifier; in the terminology of Breiman (1996), many classifiers are "unstable". Suppose we are given training data $\mathcal{X} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where each X_i is a covariate taking values in a general normed space $\mathcal{B}$, and Y_i is a response taking values in $\{-1, 1\}$. Assume further that we have access to a classifier $\hat{C}_n(\cdot) = \hat{C}_n(\cdot; \mathcal{X})$ constructed from the training data, so that $x \in \mathcal{B}$ is assigned to class $\hat{C}_n(x; \mathcal{X})$. To form the bagged version $\hat{C}_n^*$ of the classifier, we draw B bootstrap resamples $\{\mathcal{X}_b^* : b = 1, \dots, B\}$ from $\mathcal{X}$, and set

$$\hat{C}_n^*(x) := \operatorname{sgn}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{C}_n(x; \mathcal{X}_b^*)\right).$$

Peter got me interested in bagging nearest neighbour classifiers. Ironically, the nearest neighbour classifier had been described by Breiman as stable, since the nearest neighbour appears in more than half—in fact, around $1 - (1 - 1/n)^n \approx 1 - e^{-1}$—of the bootstrap resamples; thus the bagged nearest neighbour classifier is typically identical to the unbagged version. In Hall and Samworth (2005), however, we studied the effect of drawing resamples (either with or without replacement) of smaller size m. Naturally, this reduces the probability of including the nearest neighbour in the resample, and the bagged classifier is now well approximated by a weighted nearest neighbour classifier with geometrically decaying weights; see also Biau and Devroye (2010). In order for bagging to yield any asymptotic improvement over the basic nearest neighbour classifier, we require $m/n < 1/2$ (when sampling without replacement) and $m/n < \log 2$ (when sampling with replacement); in order to converge to the theoretically-optimal Bayes classifier, we require $m = m_n \to \infty$ but $m/n \to 0$.

Once classification problems had piqued his interest, Peter set about trying to answer some of the key questions on rates of convergence and tuning parameter selection that would naturally have occurred to him given his earlier work on nonparametric inference. Hall and Kang (2005) studied the performance of classifiers constructed from kernel density estimates of the class conditional distributions on $\mathcal{B} = \mathbb{R}^d$. A particularly curious discovery he made there is that even in the simplest case where $d = 1$ and where the class conditional densities f and g cross only at the single point x_0, the rate of convergence and order of the asymptotically optimal bandwidth depends on the sign of $f''(x_0)g''(x_0)$. In Hall, Park and Samworth (2008), we considered similar problems in the context of k-nearest neighbour classification, obtaining an asymptotic expansion for the regret (i.e., the difference between the risk of the k-nearest neighbour classifier and that of the Bayes classifier) which implied that the usual nonparametric error rate of order $n^{-4/(d+4)}$ was attainable with k chosen to be of order $n^{4/(d+4)}$. The form of the expansion made me realise that the limiting ratio of the regrets of the bagged nearest neighbour classifier and the k-nearest neighbour classifier (with both the resample size m and the number of neighbours k chosen optimally) depended only on d, and not on the underlying distributions. To my great surprise, this limiting ratio was greater than 1 when $d = 1$, equal to 1 when $d = 2$ and less than 1 for $d \geq 3$ (though approaching 1 for large d). It took me some years to explain this phenomenon in terms of the optimal weighting scheme [Samworth (2012)].

In more recent years, Peter turned his attention to a wealth of other important, though perhaps less well studied, issues in classification. Some of these were motivated by what he saw as drawbacks of existing classifiers. For instance, in Hall, Titterington and Xue (2009b), he developed classifiers based on componentwise medians, to alleviate the difficulties of both computing and interpreting multivariate medians; such methods can be highly effective for high-dimensional data that may have heavy tails. In Chan and Hall (2009a), he studied robust versions of nearest neighbour classifiers for high-dimensional data that try to perform an initial variable selection step to reduce variability. Chan and Hall (2009b) presented simple scale adjustments to make distance-based classifiers (primarily designed to detect location differences) less sensitive to scale variation between populations; see also Hall and Pham (2010). Hall and Xue (2010) and Hall, Xia and Xue (2013) concerned settings where one might want to incorporate the prior probabilities into a classifier, and where these prior probabilities may be significantly different from $1/2$, respectively. Finally, Ghosh and Hall (2008) discovered the phenomenon that estimating the risk of a classifier, and estimating the tuning parameters to minimise that risk, are two rather different problems, requiring the use of different methodologies.

3. Some personal reflections. I first met Peter as a Ph.D. student when he visited Cambridge in 2002. I spent an hour or so discussing a problem I was working on that involved using ideas of James–Stein estimation to find small confidence sets for the location parameter of a spherically symmetric distribution [Samworth (2005)]. I was blown away at the speed with which he was able to understand where my difficulties lay, and make helpful suggestions. Shortly afterward, he invited me to spend six weeks at the Australian National University in Canberra in July–August 2003. I arrived utterly exhausted after nearly 24 hours in the air, but Peter was full of energy when he kindly picked me up from the bus station. Almost the first thing he said to me was “I’ve got a problem I thought we could think about...”, and he proceeded to take out a pen and pad of paper; one could not help but be drawn along by his enthusiasm for research.

Everything with Peter happened at breakneck speed, whether it was dashing around the supermarket, a driving tour through the rural Australian Capital Territory (see Figure 1) or, of course, writing papers. Many of his collaborators will have experienced discussing a problem with Peter one evening and returning to the office the following morning to find that he had typed up a draft manuscript that would form the basis of a joint paper. His prose was always elegant, and he had a wonderful ability to see his way through technical asymptotic arguments, aided by almost physicist-like intuition for what ought to be true.

One of my favourite Peter stories, which I initially heard second-hand but which he later confirmed was true, concerned a time when he’d been asked to teach an elementary Statistics course to students with really very little quantitative background. Realising that he had lost some of the students along the way, and in order not to ruin their grades, Peter had a cunning idea and spent the last class before the final going through the problems that he had set on the exam. To his horror, however, the students still flunked the exam. When Peter bumped into one of the students and asked in bemusement, “What happened? I went through the questions in the last class”. The student replied, “Yes, but you did them in a different order”!

Peter had seemingly boundless energy and capacity to work, but he was also a very gentle individual in many ways. He was extraordinarily generous to others,

FIG. 1. *Peter with Juhyun Park* (*Lancaster University*), *the author and Nick Bingham* (*Imperial College London*) *on a blustery day in rural Australian Capital Territory in* 2003.

and particularly junior researchers for whom he did so much. He was a remarkable person and I miss him very deeply.

REFERENCES

BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal*. **101** 2499–2518. MR2719877

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn*. **24** 123–140.

CHAN, Y. and HALL, P. (2009a). Robust nearest-neighbor methods for classifying high-dimensional data. *Ann. Statist*. **37** 3186–3203. MR2549557

CHAN, Y.-B. and HALL, P. (2009b). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika* **96** 469–478. MR2507156

CHAN, Y. and HALL, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc*. **105** 798–809. MR2724862

CHEN, S. X. (2016). Peter Hall's contribution to the bootstrap. *Ann. Statist*. **44** 1821–1836.

CHENG, M. Y. and FAN, J. (2016). Peter Hall's contributions to nonparametric function estimation and modeling. *Ann. Statist*. **44** 1837–1853.

CHRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines*. Cambridge Univ. Press, Cambridge.

DELAIGLE, A. (2016). Peter Hall's main contributions to deconvolution. *Ann. Statist*. **44** 1854–1866.

DELAIGLE, A. and HALL, P. (2012). Effect of heavy tails on ultra high dimensional variable ranking methods. *Statist. Sinica* **22** 909–932. MR2987477

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist*. **12** 793–815. MR0751274

DÜMBGEN, L., SAMWORTH, R. J. and SCHUHMACHER, D. (2013). Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back*: *High-Dimensional Models and Processes*. *Inst. Math. Stat.* (*IMS*) *Collect*. **9** 78–90. IMS, Beachwood, OH. MR3186750

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FRIEDMAN, J. H. and HALL, P. (2007). On bagging and nonlinear estimation. *J. Statist. Plann. Inference* **137** 669–683. MR2301708

GHOSH, A. K. and HALL, P. (2008). On error-rate estimation in nonparametric classification. *Statist. Sinica* **18** 1081–1100. MR2440077

HALL, P. and KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. MR2157804

HALL, P. and LI, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889. MR1232523

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. MR2155347

HALL, P. and MILLER, H. (2009a). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18** 533–550. MR2751640

HALL, P. and MILLER, H. (2009b). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.* **37** 3929–3959. MR2572448

HALL, P., PARK, B. U. and SAMWORTH, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.* **36** 2135–2152. MR2458182

HALL, P. and PHAM, T. (2010). Optimal properties of centroid-based classifiers for very high-dimensional data. *Ann. Statist.* **38** 1071–1093. MR2604705

HALL, P. and SAMWORTH, R. J. (2005). Properties of bagged nearest neighbour classifiers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 363–379. MR2155343

HALL, P., TITTERINGTON, D. M. and XUE, J.-H. (2009a). Tilting methods for assessing the influence of components in a classifier. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 783–803. MR2750095

HALL, P., TITTERINGTON, D. M. and XUE, J.-H. (2009b). Median-based classifiers for high-dimensional data. *J. Amer. Statist. Assoc.* **104** 1597–1608. MR2597003

HALL, P., XIA, Y. and XUE, J.-H. (2013). Simple tiered classifiers. *Biometrika* **100** 431–445. MR3068444

HALL, P. and XUE, J.-H. (2010). Incorporating prior probabilities into high-dimensional classifiers. *Biometrika* **97** 31–48. MR2594415

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

MARRON, J. S., TODD, M. J. and AHN, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102** 1267–1271. MR2412548

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

MÜLLER, H.-G. (2016). Peter Hall, functional data analysis and random objects. *Ann. Statist.* **44** 1867–1887.

SAMWORTH, R. (2005). Small confidence sets for the mean of a spherically symmetric distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 343–361. MR2155342

SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618

SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE
CB3 0WB
UNITED KINGDOM
E-MAIL: r.samworth@statslab.cam.ac.uk
URL: http://www.statslab.cam.ac.uk/~rjs57