STATISTICAL SCIENCE

Volume 36, Number 2 May 2021

Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox
.... *Ruobin Gong and Xiao-Li Meng* 169
Comment: On the History and Limitations of Probability Updating .... *Glenn Shafer* 191
Comment: Settle the Unsettling: An Inferential Models Perspective
.... *Chuanhai Liu and Ryan Martin* 196
Comment: Moving Beyond Sets of Probabilities .... *Gregory Wheeler* 201
Comment: On Focusing, Soft and Strong Revision of Choquet Capacities and Their Role in Statistics .... *Thomas Augustin and Georg Schollmeyer* 205
Rejoinder: Let's Be Imprecise in Order to Be Precise (About What We Don't Know)
.... *Ruobin Gong and Xiao-Li Meng* 210
A Unified Primal Dual Active Set Algorithm for Nonconvex Sparse Recovery
.... *Jian Huang, Yuling Jiao, Bangti Jin, Jin Liu, Xiliang Lu and Can Yang* 215
The Box–Cox Transformation: Review and Extensions
.... *Anthony C. Atkinson, Marco Riani and Aldo Corbellini* 239
Noncommutative Probability and Multiplicative Cascades .... *Ian W. McKeague* 256
A Selective Overview of Deep Learning .... *Jianqing Fan, Cong Ma and Yiqiao Zhong* 264
Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications .... *Tze Leung Lai and Hongsong Yuan* 291
Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning
.... *Jianqing Fan, Kaizheng Wang, Yiqiao Zhong and Ziwei Zhu* 303
A Conversation with Dennis Cook .... *Efstathia Bura, Bing Li, Lexin Li, Christopher Nachtsheim, Daniel Pena, Claude Setodji and Robert E. Weiss* 328

***Statistical Science* [ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 36, Number 2, May 2021. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.
POSTMASTER: Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science **Volume 36, Number 2 (169–337) May 2021**

Volume 36 Number 2 May 2021

Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox
Ruobin Gong and Xiao-Li Meng

A Unified Primal Dual Active Set Algorithm for Nonconvex Sparse Recovery
Jian Huang, Yuling Jiao, Bangti Jin, Jin Liu, Xiliang Lu and Can Yang

The Box–Cox Transformation: Review and Extensions
Anthony C. Atkinson, Marco Riani and Aldo Corbellini

Noncommutative Probability and Multiplicative Cascades
Ian W. McKeague

A Selective Overview of Deep Learning
Jianqing Fan, Cong Ma and Yiqiao Zhong

Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications
Tze Leung Lai and Hongsong Yuan

Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning
Jianqing Fan, Kaizheng Wang, Yiqiao Zhong and Ziwei Zhu

A Conversation with Dennis Cook
Efstathia Bura, Bing Li, Lexin Li, Christopher Nachtsheim, Daniel Pena, Claude Setodji and Robert E. Weiss

Statistical Science
2021, Vol. 36, No. 2, 169–190
https://doi.org/10.1214/19-STS765

Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox

Ruobin Gong and Xiao-Li Meng

Abstract. Imprecise probabilities alleviate the need for high-resolution and unwarranted assumptions in statistical modeling. They present an alternative strategy to reduce irreplicable findings. However, updating imprecise models requires the user to choose among alternative updating rules. Competing rules can result in incompatible inferences, and exhibit *dilation*, *contraction* and *sure loss*, unsettling phenomena that cannot occur with precise probabilities and the regular Bayes rule. We revisit some famous statistical paradoxes and show that the logical fallacy stems from a set of marginally plausible yet jointly incommensurable model assumptions akin to the trio of phenomena above. Discrepancies between the generalized Bayes ($\mathfrak{B}$) rule, Dempster's ($\mathfrak{D}$) rule and the Geometric ($\mathfrak{G}$) rule as competing updating rules for Choquet capacities of order 2 are discussed. We note that (1) $\mathfrak{B}$-rule cannot contract nor induce sure loss, but is the most prone to dilation due to "overfitting" in a certain sense; (2) in absence of prior information, both $\mathfrak{B}$- and $\mathfrak{G}$-rules are incapable to learn from data however informative they may be; (3) $\mathfrak{D}$- and $\mathfrak{G}$-rules can mathematically contradict each other by contracting while the other dilating. These findings highlight the invaluable role of judicious judgment in handling low-resolution information, and the care that needs to be take when applying updating rules to imprecise probability models.

Key words and phrases: Imprecise probability, model uncertainty, Choquet capacity, belief function, coherence, Monty Hall problem.

REFERENCES

BALCH, M. S., MARTIN, R. and FERSON, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **475** 20180565, 20. MR3999720 https://doi.org/10.1098/rspa.2018.0565

BILLINGSLEY, P. (2013). *Convergence of Probability Measures*. Wiley, New York.

BLYTH, C. R. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* **67** 364–366, 373–381. MR0314156

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38** 325–339. MR0207001 https://doi.org/10.1214/aoms/1177698950

DIACONIS, P. (1978). Review of "A mathematical theory of evidence" (G. Shafer). *J. Amer. Statist. Assoc.* **73** 677–678.

DIACONIS, P. and ZABELL, S. (1983). Some alternatives to Bayes' Rule. Stanford University, CA. Department of Statistics.

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.

FAGIN, R. and HALPERN, J. Y. (1987). A new approach to updating beliefs. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence* 317–325.

FYGENSON, M. (2008). Modeling and predicting extrapolated probabilities with outlooks. *Statist. Sinica* **18** 9–90. MR2416904

GELMAN, A. (2006). The boxer, the wrestler, and the coin flip: A paradox of robust Bayesian inference and belief functions. *Amer. Statist.* **60** 146–150. MR2224212 https://doi.org/10.1198/000313006X106190

GONG, R. and MENG, X. L. (2021). Probabilistic underpinning of imprecise probability for statistical learning with low-resolution information. Technical Report.

GOOD, I. (1974). A little learning can be dangerous. *British J. Philos. Sci.* **25** 340–342.

HANNIG, J. and XIE, M. (2012). A note on Dempster–Shafer recombination of confidence distributions. *Electron. J. Stat.* **6** 1943–1966.

Ruobin Gong is Assistant Professor of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey, 08854, USA (e-mail: rg915@stat.rutgers.edu). Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts, 02138, USA (e-mail: meng@stat.harvard.edu).

MR2988470 https://doi.org/10.1214/12-EJS734
HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102
HEITJAN, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708. MR1326420 https://doi.org/10.1093/biomet/81.4.701
HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174 https://doi.org/10.1214/aos/1176348396
HERRON, T. SEIDENFELD, T. and WASSERMAN, L. (1994). The extent of dilation of sets of probabilities and the asymptotics of robust Bayesian inference. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 250–259.
HERRON, T., SEIDENFELD, T. and WASSERMAN, L. (1997). Divisive conditioning: Further results on dilation. *Philos. Sci.* **64** 411–444. MR1605648 https://doi.org/10.1086/392559
HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306
JAFFRAY, J.-Y. (1992). Bayesian updating and belief functions. *IEEE Trans. Syst. Man Cybern.* **22** 1144–1152. MR1202571 https://doi.org/10.1109/21.179852
KISH, L. (1965). *Survey Sampling*. Wiley, New York, NY.
KOHLAS, J. (1991). The reliability of reasoning with unreliable arguments. *Ann. Oper. Res.* **32** 67–113. MR1128173 https://doi.org/10.1007/BF02204829
KRUSE, R. and SCHWECKE, E. (1990). Specialization—a new concept for uncertainty handling with belief functions. *Int. J. Gen. Syst.* **18** 49–60.
KYBURG, H. E. (1987). Bayesian and non-Bayesian evidential updating. *Artificial Intelligence* **31** 271–293.
LIU, K. and MENG, X.-L. (2014). Comment: A fruitful resolution to Simpson's paradox via multiresolution inference. *Amer. Statist.* **68** 17–29. MR3303829 https://doi.org/10.1080/00031305.2014.876842
LIU, K. and MENG, X-L. (2016). There is individualized treatment. Why not individualized inference?. *Annu. Rev. Stat. Appl.* **3** 79–111.
MARTIN, R. and LIU, C. (2016). *Inferential Models: Reasoning with Uncertainty*. *Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727
MENG, X.-L. and XIE, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Rev.* **33** 218–250. MR3170847 https://doi.org/10.1080/07474938.2013.808567
MIRANDA, E. and MONTES, I. (2015). Coherent updating of non-additive measures. *Internat. J. Approx. Reason.* **56** 159–177. MR3278790 https://doi.org/10.1016/j.ijar.2014.05.003
MORGAN, J. P., CHAGANTY, N. R., DAHIYA, R. C. and DOVIAK, M. J. (1991). Let's make a deal: The player's dilemma. *Amer. Statist.* **45** 284–287.
MOSTELLER, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Courier Corporation, North Chelmsford, MA.
NGUYEN, H. T. (1978). On random sets and belief functions. *J. Math. Anal. Appl.* **65** 531–542.
PAVLIDES, M. G. and PERLMAN, M. D. (2009). How likely is Simpson's paradox? *Amer. Statist.* **63** 226–233. MR2750346 https://doi.org/10.1198/tast.2009.09007
PEARL, J. (1990). Reasoning with belief functions: An analysis of compatibility. *Internat. J. Approx. Reason.* **4** 5–6 363–389.
PEDERSEN, A. P. and WHEELER, G. (2014). Demystifying dilation. *Erkenntnis* **79** 1305–1342. MR3274419 https://doi.org/10.1007/s10670-013-9531-7
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.
SCHWEDER, T. and HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. *Cambridge Series in Statistical and Probabilistic Mathematics* **41**. Cambridge Univ. Press, New York. MR3558738 https://doi.org/10.1017/CBO9781139046671
SEIDENFELD, T. and WASSERMAN, L. (1993). Dilation for sets of probabilities. *Ann. Statist.* **21** 1139–1154. MR1241261 https://doi.org/10.1214/aos/1176349254
SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ.
SHAFER, G. (1979). Allocations of probability. *Ann. Probab.* **7** 827–839.
SHAFER, G. (1985). Conditional probability. International Statistical Review/Revue Internationale de Statistique 261–275.
SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.
SMETS, P. (1991). About updating. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence* 378–385.
SMETS, P. (1993). Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. *Internat. J. Approx. Reason.* **9** 1–35.
SUPPES, P. and ZANOTTI, M. (1977). On using random relations to generate upper and lower probabilities: Foundations of probability and statistics, III. *Synthese* **36** 427–440. MR0517217 https://doi.org/10.1007/BF00486106
WALLEY, P. (1981). Coherent lower (and upper) probabilities Statistics Research Report 22, University of Warwick, Coventry.
WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Taylor & Francis, Oxford, UK.
WASSERMAN, L. A. and KADANE, J. B. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.* **18** 1328–1339. MR1062711 https://doi.org/10.1214/aos/1176347752
XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496 https://doi.org/10.1111/insr.12000
YAGER, R. R. (1987). On the Dempster–Shafer framework and new combination rules. *Inform. Sci.* **41** 93–137.
YAGER, R. R. and LIU, L., eds. (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions* **219**. Springer, New York, NY.

Statistical Science
2021, Vol. 36, No. 2, 191–195
https://doi.org/10.1214/21-STS765A

Comment: On the History and Limitations of Probability Updating

Glenn Shafer

Abstract. Gong and Meng show that we can gain insights into classical paradoxes about conditional probability by acknowledging that apparently precise probabilities live within a larger world of imprecise probability. They also show that the notion of updating becomes problematic in this larger world. A closer look at the historical development of the notion of updating can give us further insights into its limitations.

Key words and phrases: Bayes's rule of conditioning, Dempster's rule, conditional probability, conditionalization, imprecise probabilities, probability protocols, relative probability, updating.

REFERENCES

ALDRICH, J. (2008). Keynes among the statisticians. *Hist. Polit. Econ.* **40** 265–316.

ALDRICH, J. (2020). Personal communication.

BERNOULLI, J. (1713). *Ars Conjectandi*. Thurnisius, Basel. See Bernoulli (2006), translation by Edith Sylla.

BERNOULLI, J. (2006). *The Art of Conjecturing. Together with "Letter to a Friend on Sets in Court Tennis*. Johns Hopkins Univ. Press, Baltimore, MD. MR2195221

BERTRAND, J. (1889). *Calcul des Probabilités*. Gauthier-Villars, Paris.

BOOLE, G. (1854). *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities*. Macmillan, London. Reprinted by Dover, New York, 1958.

BROAD, C. D. (1914). *Perception, Physics and Reality: An Enquiry into the Information That Physical Science Can Supply About the Real*. Cambridge Univ. Press, Cambridge.

BROAD, C. D. (1922). *A treatise on probability*. By J. M. Keynes. *Mind* **31** 72–85.

BRU, B. (1989). Doutes de d'Alembert sur le calcul des probabilités. In *Jean D'Alembert, Savant et Philosophe. Portrait à Plusieurs Voix* (M. Emery and P. Monzani, eds.) 279–292. Archives contemporaines, Paris.

BRU, B. (2002). Des fraises et des oranges. In *Sciences, Musiques, Lumières: Mélanges Offerts à Anne-Marie Chouillet* (U. Kölving and I. Passeron, eds.) 3–10. Centre international d'étude du XVIII^{e} siècle, Ferny-Voltaire.

COURNOT, A. A. (1843). *Exposition de la Théorie des Chances et des Probabilités*. Hachette, Paris. Reprinted in 1984 as Volume I (Bernard Bru, editor) of Cournot (1973–2010).

COURNOT, A. A. (1973–2010). *Œuvres Complètes*. Vrin, Paris. The volumes are numbered I through XI, but VI and XI are double volumes.

CZUBER, E. (1908). *Wahrscheinlichkeitsrechnung und Ihre Anwendung Auf Fehlerausgleichung, Statistik und Lebensversicherung* **1**, 2nd ed. Teubner, Leipzig.

D'ALEMBERT, J. L. R. (1767). *Éclaircissmens sur Différens Endroits des* Élémens *de Philosophie*. Chatelain, Amsterdam. In the 5th volume of d'Alembert's *Mélanges de littéraire d'histoire et de philosophie*.

DALE, A. I. (1999). *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, 2nd ed. *Sources and Studies in the History of Mathematics and Physical Sciences*. Springer, New York. MR1710390 https://doi.org/10.1007/978-1-4419-8652-8

DE FINETTI, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38** 325–339. MR0207001 https://doi.org/10.1214/aoms/1177698950

DEMPSTER, A. P. (1968). A generalization of Bayesian inference. (With discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247. MR0238428

DE MOIVRE, A. (1738). *The Doctrine of Chances: Or, a Method of Calculating the Probabilities of Events in Play*, 2nd ed. Pearson, London.

ESTES, W. K. and SUPPES, P. (1957). Foundations of Statistical Learning Theory, I. The linear model for simple learning. Technical Report No. 16. Behavioral Sciences Division, Applied Mathematics and Statistics Laboratory, Stanford Univ.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh. Subsequent editions appeared in 1959 and 1973.

GORROOCHURN, P. (2012). *Classic Problems of Probability*. Wiley, Hoboken, NJ. MR2976816 https://doi.org/10.1002/9781118314340

HAUSDORFF, F. (1901). Beiträge zur Wahrscheinlichkeitsrechnung. *Sitzungsberichte der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53** 152–178.

JEFFREYS, H. (1931). *Scientific Inference*, 1st ed. Cambridge Univ. Press, Cambridge.

JOHNSON, W. E. (1924). *Logic* **3**. Cambridge Univ. Press, Cambridge.

KEYNES, J. M. (1921). *A Treatise on Probability*. Macmillan, London.

Glenn Shafer is University Professor, Rutgers University, Newark, New Jersey (e-mail: gshafer@rutgers.edu).

KNEBEL, S. K. (2000). *Wille, Würfel und Wahrscheinlichkeit: Das System der Moralischen Notwendigkeit in der Jesuitenscholastik* 1550–1700. Meiner, Hamburg.

KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. English translation *Foundations of the Theory of Probability*, Chelsea, New York, 1950, 2nd ed. 1956.

LACROIX, S. F. (1816). *Traité Élémentaire du Calcul des Probabilités*. Courcier, Paris. Second edition 1822.

LIAGRE, J.-B.-J. (1852). *Calcul des Probabilités et Théorie des Erreurs Avec des Applications aux Sciences D'observation en Général et à la Géodésie*. Muquardt, Brussels. Second edition, 1879, prepared with the assistance of Camille Peny.

MACCOLL, H. (1896/97). The Calculus of Equivalent Statements. (Sixth Paper.). *Proc. Lond. Math. Soc.* **28** 555–579. MR1576659 https://doi.org/10.1112/plms/s1-28.1.555

MARKOV, A. A. (1900). *Probability Calculus* (*in Russian*). Imperial Academy, St. Petersburg. The second edition, which appeared in 1908, was translated into German as *Wahrscheinlichkeitsrechnung*, Teubner, Leipzig, Germany, 1912.

MCCOLL, H. (1879/80). The Calculus of Equivalent Statements. (Fourth Paper.). *Proc. Lond. Math. Soc.* **11** 113–121. MR1575250 https://doi.org/10.1112/plms/s1-11.1.113

MCCOLL, H. (1880/81). A Note on Prof. C. S. Peirce's Probability Notation of 1867. *Proc. Lond. Math. Soc.* **12** 102. MR1575784 https://doi.org/10.1112/plms/s1-12.1.102

MIROWSKI, P., ed. (1994). *Edgeworth on Chance, Economic Hazard, and Statistics*. Rowman & Littlefield, Lanham, MD.

PEIRCE, C. S. (1867). On an improvement in Boole's calculus of logic. *Proc. Amer. Acad. Arts Sci.* **7** 250–261.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. MR0464340

SHAFER, G. (1982). Bayes's two arguments for the rule of conditioning. *Ann. Statist.* **10** 1075–1089. MR0673644

SHAFER, G. (1985). Conditional probability. *Int. Stat. Rev.* **53** 261–277.

SHAFER, G. and VOVK, V. (2006). The sources of Kolmogorov's *Grundbegriffe*. *Statist. Sci.* **21** 70–98. MR2275967 https://doi.org/10.1214/088342305000000467

TELLER, P. (1973). Conditionalization and observation. *Synthese* **26** 218–258.

WRINCH, D. and JEFFREYS, H. (1919). On some aspects of the theory of probability. *The London, Edinburgh and Dublin Philosophical Magazine, Series 6* **38** 715–731.

Statistical Science
2021, Vol. 36, No. 2, 196–200
https://doi.org/10.1214/21-STS765B

Comment: Settle the Unsettling: An Inferential Models Perspective

Chuanhai Liu and Ryan Martin

Abstract. Here, we demonstrate that the *inferential model* (IM) framework, unlike the updating rules that Gong and Meng show to be unreliable, provides valid and efficient inferences/prediction while not being susceptible to sure loss. In this sense, the IM framework settles what Gong and Meng characterized as "unsettling."

Key words and phrases: Belief function, efficiency, lower and upper probability, inferential models, validity.

REFERENCES

CELLA, L. and MARTIN, R. (2020). Validity, consonant plausibility measures, and conformal prediction. Available at https://researchers.one/articles/20.01.00010.

DEMPSTER, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.* **48** 365–377. MR2419025 https://doi.org/10.1016/j.ijar.2007.03.004

GONG, R. and MENG, X.-L. (2021). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statist. Sci.* **36** 169–190.

MARTIN, R. and LINGHAM, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics* **58** 225–235. MR3488301 https://doi.org/10.1080/00401706.2015.1017116

MARTIN, R. and LIU, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.* **108** 301–313. MR3174621 https://doi.org/10.1080/01621459.2012.747960

MARTIN, R. and LIU, C. (2015a). Conditional inferential models: Combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 195–217. MR3299405 https://doi.org/10.1111/rssb.12070

MARTIN, R. and LIU, C. (2015b). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.* **110** 1621–1631. MR3449059 https://doi.org/10.1080/01621459.2014.985827

MARTIN, R. and LIU, C. (2016). *Inferential Models: Reasoning with Uncertainty. Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. MR0464340

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities. Monographs on Statistics and Applied Probability* **42**. CRC Press, London. MR1145491 https://doi.org/10.1007/978-1-4899-3472-7

Chuanhai Liu is Professor, Department of Statistics, Purdue University, West Lafayette, Indiana, USA (e-mail: chuanhai@purdue.edu). Ryan Martin is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA (e-mail: rgmarti3@ncsu.edu).

Statistical Science
2021, Vol. 36, No. 2, 201–204
https://doi.org/10.1214/21-STS765C

Comment: Moving Beyond Sets of Probabilities

Gregory Wheeler

Abstract. The theory of lower previsions is designed around the principles of coherence and sure-loss avoidance, thus steers clear of all the updating anomalies highlighted in Gong and Meng's "Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox" except dilation. In fact, the traditional problem with the theory of imprecise probability is that coherent inference is too complicated rather than unsettling. Progress has been made simplifying coherent inference by demoting sets of probabilities from fundamental building blocks to secondary representations that are derived or discarded as needed.

Key words and phrases: Desirable gambles, lower previsions, imprecise probabilities, dilation.

REFERENCES

COUSO, I., MORAL, S. and WALLEY, P. (1999). Examples of independence for imprecise probabilities. In *Proceedings of the First Symposium on Imprecise Probabilities and Their Applications* (*ISIPTA*) (G. de Cooman, ed.). Ghent, Belgium.

DE BOCK, J. and DE COOMAN, G. (2015). Conditioning, updating and lower probability zero. *Internat. J. Approx. Reason.* **67** 1–36. MR3415761 https://doi.org/10.1016/j.ijar.2015.08.006

DE BOCK, J. and DE COOMAN, G. (2019). Interpreting, axiomatising and representing coherent choice functions in terms of desirability. In *Proceedings of the Eleventh International Symposium on Imprecise Probabilities*: *Theories and Applications* (J. De Bock, C. P. de Campos, G. de Cooman, E. Quaeghebeur and G. Wheeler, eds.). *Proc. Mach. Learn. Res.* (*PMLR*) **103** 125–134. Thagaste, Ghent, Belgium. MR4020913

DE FINETTI, B. and SAVAGE, L. J. (1962). Sul modo di scegliere le probabilità iniziali. *Bibl. Metron, Ser. C* **1** 81–154.

GONG, R. and MENG, X.-L. (2021). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statist. Sci.*

PEDERSEN, A. P. and WHEELER, G. (2014). Demystifying dilation. *Erkenntnis* **79** 1305–1342. MR3274419 https://doi.org/10.1007/s10670-013-9531-7

PEDERSEN, A. P. and WHEELER, G. (2015). Dilation, disintegrations, and delayed decisions. In *Proceedings of the* 9*th Symposium on Imprecise Probabilities and Their Applications* (*ISIPTA*) 227–236, Pescara, Italy.

PEDERSEN, A. P. and WHEELER, G. (2019). Dilation and asymmetric relevance. In *Proceedings of the* 11*th Symposium on Imprecise Probabilities and Their Applications* (*ISIPTA*) (J. De Bock, C. P. Campos, G. de Cooman, E. Quaeghebeur and G. Wheeler, eds.). *Proc. Mach. Learn. Res.* (*PMLR*) **103** 324–326.

SMITH, C. A. B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. Ser. B* **23** 1–37. MR0124097

TROFFAES, M. C. M. and DE COOMAN, G. (2014). *Lower Previsions*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3222242 https://doi.org/10.1002/9781118762622

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. *Monographs on Statistics and Applied Probability* **42**. CRC Press, London. MR1145491 https://doi.org/10.1007/978-1-4899-3472-7

WHEELER, G. (2021). A gentle approach to imprecise probability. In *Reflections on the Foundations of Statistics*: *Essays in Honor of Teddy Seidenfeld* (T. Augustin, F. Cozman and G. Wheeler, eds.). *Theory and Decision Library A*. Springer, Berlin.

Gregory Wheeler is Professor of Theoretical Philosophy and Computer Science, Frankfurt School of Finance & Management, 60322 Frankfurt am Main, Germany (e-mail: g.wheeler@fs.de).

Statistical Science
2021, Vol. 36, No. 2, 205–209
https://doi.org/10.1214/21-STS765D

Comment: On Focusing, Soft and Strong Revision of Choquet Capacities and Their Role in Statistics

Thomas Augustin and Georg Schollmeyer

Abstract. We congratulate Ruobin Gong and Xiao-Li Meng on their thought-provoking paper demonstrating the power of imprecise probabilities in statistics. In particular, Gong and Meng clarify important statistical paradoxes by discussing them in the framework of generalized uncertainty quantification and different conditioning rules used for updating. In this note, we characterize all three conditioning rules as envelopes of certain sets of conditional probabilities. This view also suggests some generalizations that can be seen as compromise rules. Similar to Gong and Meng, our derivations mainly focus on Choquet capacities of order 2, and so we also briefly discuss in general their role as statistical models. We conclude with some general remarks on the potential of imprecise probabilities to cope with the multidimensional nature of uncertainty.

Key words and phrases: Imprecise probabilities, Choquet capacities, updating, neighborhood models, generalized Bayes rule, Dempster's rule of conditioning.

REFERENCES

AUGUSTIN, T. and HABLE, R. (2010). On the impact of robust statistics on imprecise probability models: A review. *Struct. Saf.* **32** 358–365.

AUGUSTIN, T., WALTER, G. and COOLEN, F. P. A. (2014). Statistical inference. In *Introduction to Imprecise Probabilities*. *Wiley Ser. Probab. Stat.* 135–189. Wiley, Chichester. MR3287404 https://doi.org/10.1002/9781118763117.ch7

BENAVOLI, A. and ZAFFALON, M. (2015). Prior near ignorance for inferences in the k-parameter exponential family. *Statistics* **49** 1104–1140. MR3378034 https://doi.org/10.1080/02331888.2014.960869

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

CATTANEO, M. E. G. V. (2014). A continuous updating rule for imprecise probabilities. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. *Part III*. *Commun. Comput. Inf. Sci.* **444** 426–435. Springer, Cham. MR3616468

CHATEAUNEUF, A. and JAFFRAY, J.-Y. (1989). Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Math. Social Sci.* **17** 263–283. MR1006179 https://doi.org/10.1016/0165-4896(89)90056-5

DENŒUX, T. (2016). 40 years of Dempster–Shafer theory [editorial]. *Internat. J. Approx. Reason.* **79** 1–6. MR3548391 https://doi.org/10.1016/j.ijar.2016.07.010

DESTERCKE, S., DUBOIS, D. and CHOJNACKI, E. (2008). Unifying practical uncertainty representations. I. Generalized p-boxes. *Internat. J. Approx. Reason.* **49** 649–663. MR2475260 https://doi.org/10.1016/j.ijar.2008.07.003

DUBOIS, D. and PRADE, H. (1997). Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge. In *Qualitative and Quantitative Practical Reasoning* (D. M. Gabbay, R. Kruse, A. Nonnengart and H. J. Ohlbach, eds.) 96–107. Springer, Berlin, Heidelberg.

FIERENS, P. I., RÊGO, L. C. and FINE, T. L. (2009). A frequentist understanding of sets of measures. *J. Statist. Plann. Inference* **139** 1879–1892. MR2497546 https://doi.org/10.1016/j.jspi.2008.08.025

GILBOA, I. and SCHMEIDLER, D. (1993). Updating ambiguous beliefs. *J. Econom. Theory* **59** 33–49. MR1211549 https://doi.org/10.1006/jeth.1993.1003

Thomas Augustin is Professor of Statistics and Head of the Foundations of Statistics and their Applications Group, Department of Statistics, Ludwig-Maximilians Universität München (LMU Munich), Ludwigstr. 33, D-80539 Munich, Germany (e-mail: thomas.augustin@stat.uni-muenchen.de). Georg Schollmeyer is a post-doctorial staff member of the Foundations of Statistics and their Applications Group, Department of Statistics, Ludwig-Maximilians Universität München (LMU Munich), Ludwigstr. 33, D-80539 Munich, Germany (e-mail: georg.schollmeyer@stat.uni-muenchen.de).

GILBOA, I. and SCHMEIDLER, D. (1994). Additive representations of non-additive measures and the Choquet integral. *Ann. Oper. Res.* **52** 43–65. MR1293559 https://doi.org/10.1007/BF02032160

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. Univ. Minnesota Press, Minneapolis, MN. MR0723501

HELD, H., AUGUSTIN, T. and KRIEGLER, E. (2008). Bayesian learning for a class of priors with prescribed marginals. *Internat. J. Approx. Reason.* **49** 212–233. MR2454840 https://doi.org/10.1016/j.ijar.2008.03.018

HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306

KLIR, G. J. and WIERMAN, M. J. (1999). *Uncertainty-Based Information. Elements of Generalized Information Theory*, 2nd ed. *Studies in Fuzziness and Soft Computing* **15**. Physica-Verlag, Heidelberg. MR1719879 https://doi.org/10.1007/978-3-7908-1869-7

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions. Springer Series in Statistics*. Springer, New York. MR2151380

MOLCHANOV, I. and MOLINARI, F. (2018). *Random Sets in Econometrics. Econometric Society Monographs* **60**. Cambridge Univ. Press, Cambridge. MR3753715 https://doi.org/10.1017/9781316392973

MOLINARI, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics. Volume* 7, *Part A* (S. N. Durlauf, L. P. Hansen, J. J. Heckman and R. L. Matzkin, eds.) 355–486.

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020a). Unifying neighbourhood and distortion models: Part I—new results on old models. *Int. J. Gen. Syst.* **49** 602–635. MR4151121 https://doi.org/10.1080/03081079.2020.1778682

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020b). Unifying neighbourhood and distortion models: Part II—new models and synthesis. *Int. J. Gen. Syst.* **49** 636–674. MR4151122 https://doi.org/10.1080/03081079.2020.1778683

SHAFER, G. (1990). Perspectives on the theory and practice of belief functions. *Internat. J. Approx. Reason.* **4** 323–362. MR1078973 https://doi.org/10.1016/0888-613X(90)90012-Q

WALLEY, P. (1981). Coherent lower (and upper) probabilities Technical Report Univ. Warwick Coventry. Statistics Research Report.

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities. Monographs on Statistics and Applied Probability* **42**. CRC Press, London. MR1145491 https://doi.org/10.1007/978-1-4899-3472-7

WALLEY, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58** 3–57. MR1379233

WANG, S. S. and YOUNG, V. R. (1998). Risk-adjusted credibility premiums using distorted probabilities. *Scand. Actuar. J.* **2** 143–165. MR1659301 https://doi.org/10.1080/03461238.1998.10413999

WEICHSELBERGER, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept* [*Elementary Foundations of a More General Calculus of Probability I: Interval Probability as a Comprehensive Concept*; *in German*]. Physica, Heidelberg.

WEICHSELBERGER, K. and PÖHLMANN, S. (1990). *A Methodology for Uncertainty in Knowledge-Based Systems. Lecture Notes in Computer Science* **419**. Springer, Berlin. MR1048906

Statistical Science
2021, Vol. 36, No. 2, 210–214
https://doi.org/10.1214/21-STS765REJ

Rejoinder: Let's Be Imprecise in Order to Be Precise (About What We Don't Know)

Ruobin Gong and Xiao-Li Meng

REFERENCES

CARNAP, R. (1962). *Logic, Methodology and Philosophy of Science. Proceedings of the* 1960 *International Congress*. Stanford Univ. Press, Stanford, CA. MR0166069

COPAS, J. B. (1969). Compound decisions and empirical Bayes. *J. Roy. Statist. Soc. Ser. B* **31** 397–417. MR0269013

DECADT, A., DE COOMAN, G. and DE BOCK, J. (2019). Monte Carlo estimation for imprecise probabilities: Basic properties. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 135–144. MR4020914

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Stat.* **37** 355–374. MR0187357 https://doi.org/10.1214/aoms/1177699517

DEMPSTER, A. P. (1972). A class of random convex polytopes. *Ann. Math. Stat.* **43** 260–272. MR0303650 https://doi.org/10.1214/aoms/1177692719

FETZ, T. (2019). Improving the convergence of iterative importance sampling for computing upper and lower expectations. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 185–193. MR4020919

GONG, R. and MENG, X.-L. (2021). Probabilistic underpinning of imprecise probability for statistical learning with low-resolution information. Technical report.

GOOD, I. J. (1966). Symposium on current views of subjective probability: Subjective probability as the measure of a non-measurable set. In *Studies in Logic and the Foundations of Mathematics* **44** 319–329. Elsevier, Amsterdam.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B* **29** 399–418. MR0225428

JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules. Preprint. Available at arXiv:1708.08719.

JACOB, P. E., GONG, R., EDLEFSEN, P. T. and DEMPSTER, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *J. Amer. Statist. Assoc.* To appear. With discussion.

LI, X. and MENG, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1844210

LINDLEY, D. V. (1969). Discussion of compound decisions and empirical Bayes, J. B. Copas. *J. Roy. Statist. Soc. Ser. B* **31** 419–421.

LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241 https://doi.org/10.1214/09-BA404

LIU, K. and MENG, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annu. Rev. Stat. Appl.* **3** 79–111.

LUNN, D., BEST, N., SPIEGELHALTER, D., GRAHAM, G. and NEUENSCHWANDER, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *J. Pharmacokinet. Pharmacodyn.* **36** 19.

MARTIN, R. and LIU, C. (2015). *Inferential Models: Reasoning with Uncertainty*. CRC Press, Boca Raton, FL.

MARTIN, R. and SYRING, N. (2019). Validity-preservation properties of rules for combining inferential models. In *International Symposium on Imprecise Probabilities: Theories and Applications. Proc. Mach. Learn. Res.* (*PMLR*) **103** 286–294. MR4020929

MENG, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science* (X. Lin et al., eds.) 537–562.

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12** 685–726. MR3834282 https://doi.org/10.1214/18-AOAS1161SF

PLUMMER, M. (2015). Cuts in Bayesian graphical models. *Stat. Comput.* **25** 37–43. MR3304902 https://doi.org/10.1007/s11222-014-9503-z

TELLER, P. (1973). Conditionalization and observation. *Synthese* **26** 218–258.

Ruobin Gong is Assistant Professor of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854, USA (e-mail: rg915@stat.rutgers.edu). Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: meng@stat.harvard.edu).

Statistical Science
2021, Vol. 36, No. 2, 215–238
https://doi.org/10.1214/19-STS758

A Unified Primal Dual Active Set Algorithm for Nonconvex Sparse Recovery

Jian Huang, Yuling Jiao, Bangti Jin, Jin Liu, Xiliang Lu and Can Yang

Abstract. In this paper, we consider the problem of recovering a sparse signal based on penalized least squares formulations. We develop a novel algorithm of primal-dual active set type for a class of nonconvex sparsity-promoting penalties, including ℓ^0, bridge, smoothly clipped absolute deviation, capped ℓ^1 and minimax concavity penalty. First, we establish the existence of a global minimizer for the related optimization problems. Then we derive a novel necessary optimality condition for the global minimizer using the associated thresholding operator. The solutions to the optimality system are coordinatewise minimizers, and under minor conditions, they are also local minimizers. Upon introducing the dual variable, the active set can be determined using the primal and dual variables together. Further, this relation lends itself to an iterative algorithm of active set type which at each step involves first updating the primal variable only on the active set and then updating the dual variable explicitly. When combined with a continuation strategy on the regularization parameter, the primal dual active set method is shown to converge globally to the underlying regression target under certain regularity conditions. Extensive numerical experiments with both simulated and real data demonstrate its superior performance in terms of computational efficiency and recovery accuracy compared with the existing sparse recovery methods.

Key words and phrases: Nonconvex penalty, sparsity, primal-dual active set algorithm, continuation, consistency.

REFERENCES

[1] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

[2] BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

[3] BLUMENSATH, T. and DAVIES, M. E. (2008). Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.* **14** 629–654. MR2461601 https://doi.org/10.1007/s00041-008-9035-z

[4] BREDIES, K., LORENZ, D. A. and REITERER, S. (2015). Minimization of non-smooth, non-convex functionals by iterative thresholding. *J. Optim. Theory Appl.* **165** 78–112. MR3327417 https://doi.org/10.1007/s10957-014-0614-7

[5] BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396 https://doi.org/10.1214/10-AOAS388

[6] BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957 https://doi.org/10.1214/aos/1032181158

[7] CANDÈS, E. J., ROMBERG, J. and TAO, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52** 489–509. MR2236170 https://doi.org/10.1109/TIT.2005.862083

[8] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

Jian Huang is Professor, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242 (e-mail: jian-huang@uiowa.edu). Yuling Jiao is Associate Professor, School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, 430063, P.R. China (e-mail: yulingjiaomath@whu.edu.cn). Bangti Jin is Professor, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (e-mail: b.jin@ucl.ac.uk; bangti.jin@gmail.com). Jin Liu is Assistant Professor, Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore (e-mail: jin.liu@duke-nus.edu.sg). Xiliang Lu is Professor, School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China, and Hubei Key Laboratory of Computational Science (Wuhan University), Wuhan, 430072, China (e-mail: xllv.math@whu.edu.cn). Can Yang is Associate Professor, Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (e-mail: macyang@ust.hk).

[9] CHARTRAND, R. and STANEVA, V. (2008). Restricted isometry properties and nonconvex compressive sensing. *Inverse Probl*. **24** 035020, 14. MR2421974 https://doi.org/10.1088/0266-5611/24/3/035020

[10] CHARTRAND, R. and YIN, W. (2008). Iteratively reweighted algorithms for compressive sensing. In Proc. ICASSP 2008 3869–3872.

[11] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput*. **20** 33–61. MR1639094 https://doi.org/10.1137/S1064827596304010

[12] CHEN, X. (2012). Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program*. **134** 71–99. MR2947553 https://doi.org/10.1007/s10107-012-0569-0

[13] CHEN, X., NASHED, Z. and QI, L. (2000). Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal*. **38** 1200–1216. MR1786137 https://doi.org/10.1137/S0036142999356719

[14] CHEN, X., NIU, L. and YUAN, Y. (2013). Optimality conditions and a smoothing trust region Newton method for nonLipschitz optimization. *SIAM J. Optim*. **23** 1528–1552. MR3082500 https://doi.org/10.1137/120871390

[15] CHEN, Y., GE, D., WANG, M., WANG, Z., YE, Y. and YIN, H. (2017). Strong NP-hardness for sparse optimization with concave penalty functions. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 740–747.

[16] COMBETTES, P. L. and WAJS, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul*. **4** 1168–1200. MR2203849 https://doi.org/10.1137/050626090

[17] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory* **52** 1289–1306. MR2241189 https://doi.org/10.1109/TIT.2006.871582

[18] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc*. **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[19] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist*. **32** 928–961. MR2065194 https://doi.org/10.1214/009053604000000256

[20] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist*. **42** 819–849. MR3210988 https://doi.org/10.1214/13-AOS1198

[21] FAN, Q., JIAO, Y. and LU, X. (2014). A primal dual active set algorithm with continuation for compressed sensing. *IEEE Trans. Signal Process*. **62** 6276–6285. MR3281518 https://doi.org/10.1109/TSP.2014.2362880

[22] FENG, L. and ZHANG, C.-H. (2019). Sorted concave penalized regression. *Ann. Statist*. **47** 3069–3098. MR4025735 https://doi.org/10.1214/18-AOS1759

[23] FOUCART, S. and LAI, M.-J. (2009). Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal*. **26** 395–407. MR2503311 https://doi.org/10.1016/j.acha.2008.09.001

[24] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

[25] FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist*. **7** 397–416. MR1646710 https://doi.org/10.2307/1390712

[26] GASSO, G., RAKOTOMAMONJY, A. and CANU, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. Signal Process*. **57** 4686–4698. MR2722328 https://doi.org/10.1109/TSP.2009.2026004

[27] GELFAND, I. M. and FOMIN, S. V. (1963). *Calculus of Variations. Revised English Edition Translated and Edited by Richard A. Silverman*. Prentice-Hall, Inc., Englewood Cliffs, N.J. MR0160139

[28] GONG, P., ZHANG, C., LU, Z., HUANG, J. and YE, J. (2013). A general iterative shrinkage and thresholding algorithm for nonconvex regularized optimization problems. In *Proc*. 30*th Int. Conf. Mach. Learn* 37–45.

[29] GRIESSE, R. and LORENZ, D. A. (2008). A semismooth Newton method for Tikhonov functionals with sparsity constraints. *Inverse Probl*. **24** 035007, 19. MR2421961 https://doi.org/10.1088/0266-5611/24/3/035007

[30] HAN, Z. (2015). Primal-Dual Active-Set Methods for Convex Quadratic Optimization with Applications Ph.D. thesis, Lehigh University, PA. Available at http://coral.ie.lehigh.edu/~pubs/files/Zheng_Han.pdf.

[31] HINTERMÜLLER, M., ITO, K. and KUNISCH, K. (2002). The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim*. **13** 865–888 (2003). MR1972219 https://doi.org/10.1137/S1052623401383558

[32] HINTERMÜLLER, M. and WU, T. (2014). A superlinearly convergent R-regularized Newton scheme for variational models with concave sparsity-promoting priors. *Comput. Optim. Appl*. **57** 1–25. MR3146498 https://doi.org/10.1007/s10589-013-9583-2

[33] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist*. **36** 587–613. MR2396808 https://doi.org/10.1214/009053607000000875

[34] HUANG, J., JIAO, Y., LIU, Y. and LU, X. (2018). A constructive approach to L_0 penalized regression. *J. Mach. Learn. Res*. **19** Paper No. 10, 37. MR3862417

[35] HUANG, J., JIAO, Y., LU, X., SHI, Y. and YANG, Q. (2018). SNAP: a semismooth Newton algorithm for pathwise optimization with optimal local convergence rate and oracle properties. Preprint. Available at arXiv:1810.03814.

[36] HUANG, J., JIAO, Y., LU, X. and ZHU, L. (2018). Robust decoding from 1-bit compressive sampling with ordinary and regularized least squares. *SIAM J. Sci. Comput*. **40** A2062–A2086. MR3820388 https://doi.org/10.1137/17M1154102

[37] HUANG, S. and ZHU, J. (2011). Recovery of sparse signals using OMP and its variants: Convergence analysis based on RIP. *Inverse Probl*. **27** 035003, 14. MR2772522 https://doi.org/10.1088/0266-5611/27/3/035003

[38] ITO, K. and JIN, B. (2015). *Inverse Problems: Tikhonov Theory and Algorithms. Series on Applied Mathematics* **22**. World Scientific, Hackensack, NJ. MR3244283

[39] ITO, K. and KUNISCH, K. (2008). *Lagrange Multiplier Approach to Variational Problems and Applications. Advances in Design and Control* **15**. SIAM, Philadelphia, PA. MR2441683 https://doi.org/10.1137/1.9780898718614

[40] ITO, K. and KUNISCH, K. (2014). A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse Probl*. **30** 015001, 23. MR3151680 https://doi.org/10.1088/0266-5611/30/1/015001

[41] JIAO, Y., JIN, B. and LU, X. (2015). A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem. *Appl. Comput. Harmon. Anal*. **39** 400–426. MR3398943 https://doi.org/10.1016/j.acha.2014.10.001

[42] JOST, J. and LI-JOST, X. (1998). *Calculus of Variations. Cambridge Studies in Advanced Mathematics* **64**. Cambridge Univ. Press, Cambridge. MR1674720

[43] KINGSBURY, N. (2001). Complex wavelets for shift invariant analysis and filtering of signals. *Appl. Comput. Harmon. Anal*. **10** 234–253. MR1829802 https://doi.org/10.1006/acha.2000.0343

[44] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787 https://doi.org/10.1214/aos/1015957397

[45] KUMMER, B. (1988). Newton's method for nondifferentiable functions. In *Advances in Mathematical Optimization. Math. Res.* **45** 114–125. Akademie-Verlag, Berlin. MR0953328

[46] LAI, M.-J. and WANG, J. (2011). An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. *SIAM J. Optim.* **21** 82–101. MR2765490 https://doi.org/10.1137/090775397

[47] LAI, M.-J., XU, Y. and YIN, W. (2013). Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J. Numer. Anal.* **51** 927–957. MR3033038 https://doi.org/10.1137/110840364

[48] LI, X., SUN, D. and TOH, K.-C. (2018). A highly efficient semismooth Newton augmented Lagrangian method for solving lasso problems. *SIAM J. Optim.* **28** 433–458. MR3763098 https://doi.org/10.1137/16M1097572

[49] LIU, H., YAO, T. and LI, R. (2016). Global solutions to folded concave penalized nonconvex learning. *Ann. Statist.* **44** 629–659. MR3476612 https://doi.org/10.1214/15-AOS1380

[50] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[51] LU, Z. (2014). Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Math. Program.* **147** 277–307. MR3258526 https://doi.org/10.1007/s10107-013-0722-4

[52] MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769 https://doi.org/10.1198/jasa.2011.tm09738

[53] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[54] MO, Q. and SHEN, Y. (2012). A remark on the restricted isometry property in orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **58** 3654–3656. MR2924391 https://doi.org/10.1109/TIT.2012.2185923

[55] NEEDELL, D. and TROPP, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26** 301–321. MR2502366 https://doi.org/10.1016/j.acha.2008.07.002

[56] NIKOLOVA, M. (2013). Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer. *SIAM J. Imaging Sci.* **6** 904–937. MR3055237 https://doi.org/10.1137/11085476X

[57] PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575.

[58] SABATTI, C., HARTIKAINEN, A.-L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILO, T., KAAKINEN, M. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41** 35–46.

[59] SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.* **3** 384–415. MR2501318 https://doi.org/10.1214/08-EJS348

[60] SUN, Q. (2012). Recovery of sparsest signals via ℓ^q-minimization. *Appl. Comput. Harmon. Anal.* **32** 329–341. MR2892737 https://doi.org/10.1016/j.acha.2011.07.001

[61] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[62] TROPP, J. A. and GILBERT, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53** 4655–4666. MR2446929 https://doi.org/10.1109/TIT.2007.909108

[63] TROPP, J. A. and WRIGHT, S. J. (2010). Computational methods for sparse solution of linear inverse problems. *Proc. IEEE* **98** 948–958.

[64] TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069 https://doi.org/10.1023/A:1017501703105

[65] WANG, J., WANG, L. J., ZHONG, Y., GU, P., SHAO, J. Q., JIANG, S. S. and GONG, J. B. (2013). CETP gene polymorphisms and risk of coronary atherosclerosis in a Chinese population. *Lipids Health Dis.* **12** 176.

[66] WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873 https://doi.org/10.1214/13-AOS1159

[67] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977 https://doi.org/10.1214/14-AOS1238

[68] WILLER, C. J., SANNA, S., JACKSON, A. U., SCUTERI, A., BONNYCASTLE, L. L., CLARKE, R., HEATH, S. C., TIMPSON, N. J., NAJJAR, S. S. et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40** 161–169.

[69] ZEMUNIK, T., BOBAN, M., LAUC, G., JANKOVIC, S., ROTIM, K., VATAVUK, Z., BENCIC, G., DOGAS, Z., BORASKA, V. et al. (2009). Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat. Med. J.* **50** 23–33.

[70] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

[71] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

[72] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135 https://doi.org/10.1214/12-STS399

[73] ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[74] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[75] ZHAO, Y.-B. and LI, D. (2012). Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems. *SIAM J. Optim.* **22** 1065–1088. MR3023764 https://doi.org/10.1137/110847445

[76] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

Statistical Science
2021, Vol. 36, No. 2, 239–255
https://doi.org/10.1214/20-STS778

The Box–Cox Transformation: Review and Extensions

Anthony C. Atkinson, Marco Riani and Aldo Corbellini

Abstract. The Box–Cox power transformation family for nonnegative responses in linear models has a long and interesting history in both statistical practice and theory, which we summarize. The relationship between generalized linear models and log transformed data is illustrated. Extensions investigated include the transform both sides model and the Yeo–Johnson transformation for observations that can be positive or negative. The paper also describes an extended Yeo–Johnson transformation that allows positive and negative responses to have different power transformations. Analyses of data show this to be necessary. Robustness enters in the fan plot for which the forward search provides an ordering of the data. Plausible transformations are checked with an extended fan plot. These procedures are used to compare parametric power transformations with nonparametric transformations produced by smoothing.

Key words and phrases: ACE, AVAS, constructed variable, extended Yeo–Johnson transformation, forward search, linked plots, robust methods.

REFERENCES

ANDREWS, D. F., GNANADESIKAN, R. and WARNER, J. L. (1971). Transformations of multivariate data. *Biometrics* **27** 825–840.

ARANDA-ORDAZ, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68** 357–363. MR0626394 https://doi.org/10.1093/biomet/68.2.357

ATKINSON, A. C. (1973). Testing transformations to normality. *J. Roy. Statist. Soc. Ser. B* **35** 473–479. MR0345274

ATKINSON, A. C. (1982). Regression diagnostics, transformations and constructed variables. *J. Roy. Statist. Soc. Ser. B* **44** 1–36. MR0655369

ATKINSON, A. C. (1985). *Plots, Transformations, and Regression*. Oxford Univ. Press, Oxford.

ATKINSON, A. C. and COX, D. R. (1988). Transformations I. *Wiley Statsref.*

ATKINSON, A. C., KOOPMAN, S. J. and SHEPHARD, N. (1997). Detecting shocks: Outliers and breaks in time series. *J. Econometrics* **80** 387–422. MR1483186 https://doi.org/10.1016/S0304-4076(97)00050-X

ATKINSON, A. C. and LAWRANCE, A. J. (1989). A comparison of asymptotically equivalent test statistics for regression transformation. *Biometrika* **76** 223–229. MR1016012 https://doi.org/10.1093/biomet/76.2.223

ATKINSON, A. C., PERICCHI, L. R. and SMITH, R. L. (1991). Grouped likelihood for the shifted power transformation. *J. Roy. Statist. Soc. Ser. B* **53** 473–482. MR1108342

ATKINSON, A. and RIANI, M. (2000). *Robust Diagnostic Regression Analysis*. *Springer Series in Statistics*. Springer, New York. MR1884997 https://doi.org/10.1007/978-1-4612-1160-0

ATKINSON, A. C. and RIANI, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemom. Intell. Lab. Syst.* **60** 87–100.

ATKINSON, A. C., RIANI, M. and CERIOLI, A. (2004). *Exploring Multivariate Data with the Forward Search*. *Springer Series in Statistics*. Springer, New York. MR2055967 https://doi.org/10.1007/978-0-387-21840-3

ATKINSON, A. C., RIANI, M. and CERIOLI, A. (2010). The forward search: Theory and data analysis. *J. Korean Statist. Soc.* **39** 117–134. MR2758131 https://doi.org/10.1016/j.jkss.2010.02.007

ATKINSON, A. C., RIANI, M. and CORBELLINI, A. (2020). The transformation of profit and loss data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 251–275. https://doi.org/10.1111/rssc.12389

ATKINSON, A. C., RIANI, M. and CORBELLINI, A. (2021). Supplement to "The Box–Cox transformation: Review and extensions." https://doi.org/10.1214/20-STS778SUPP

BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311. MR0624332

BLISS, C. I. (1934). The method of probits. *Science* **79** 38–39.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BOX, G. E. P. and COX, D. R. (1982). Comment on: "An analysis of transformations revisited" [J. Amer. Statist. Assoc. **76** (1981), no. 374, 296–311; MR 83b:62048] by P. J. Bickel and K. A. Doksum. *J. Amer. Statist. Assoc.* **77** 209–210. MR0648047

Anthony C. Atkinson is Emeritus Professor, Department of Statistics, London School of Economics, London WC2A 2AE, United Kingdom (e-mail: a.c.atkinson@lse.ac.uk). Marco Riani is Full Professor, Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma, Italy (e-mail: mriani@unipr.it). Aldo Corbellini is Senior Lecturer, Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, 43100 Parma, Italy (e-mail: aldo.corbellini@unipr.it).

BOX, G. E. P. and TIDWELL, P. W. (1962). Transformation of the independent variables. *Technometrics* **4** 531–550. MR0184313 https://doi.org/10.2307/1266288

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

BREUSCH, T. S. and PAGAN, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47** 1287–1294. MR0545960 https://doi.org/10.2307/1911963

BROWN, B. W. and HOLLANDER, M. (1977). *Statistics*: *A Biomedical Introduction*. Wiley, New York.

BUJA, A. and KASS, R. E. (1985). Comment on "Estimating optimal transformations for multiple regression and transformation" by Breiman and Friedman. *J. Amer. Statist. Assoc.* **80** 602–607.

CAI, T., TIAN, L. and WEI, L. J. (2005). Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika* **92** 619–632. MR2202650 https://doi.org/10.1093/biomet/92.3.619

CARROLL, R. J. (1982). Prediction and power transformations when the choice of power is restricted to a finite set. *J. Amer. Statist. Assoc.* **77** 908–915. MR0686417

CARROLL, R. J. and RUPPERT, D. (1985). Transformations in regression: A robust analysis. *Technometrics* **27** 1–12. MR0772893 https://doi.org/10.2307/1270463

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. *Monographs on Statistics and Applied Probability*. CRC Press, New York. MR1014890 https://doi.org/10.1007/978-1-4899-2873-3

CHEN, G., LOCKHART, R. A. and STEPHENS, M. A. (2002). Box–Cox transformations in linear models: Large sample theory and tests of normality. *Canad. J. Statist.* **30** 177–234. MR1926062 https://doi.org/10.2307/3315946

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0675263

COX, D. R. (1977). Nonlinear models, residuals and transformations. *Math. Operationsforsch. Stat. Ser. Stat.* **8** 3–22. MR0488387 https://doi.org/10.1080/02331887708801354

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. MR0893334

FIRTH, D. (1988). Multiplicative errors: Log-normal or gamma? *J. Roy. Statist. Soc. Ser. B* **50** 266–268. MR0964179

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

FOSTER, A. M., TIAN, L. and WEI, L. J. (2001). Estimation for the Box–Cox transformation model without assuming parametric error distribution. *J. Amer. Statist. Assoc.* **96** 1097–1101. MR1947257 https://doi.org/10.1198/016214501753208753

FRIEDMAN, J. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report ORION 003, Department of Statistics, Stanford Univ.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York. MR0440802

GOTTARDO, R. and RAFTERY, A. (2009). Bayesian robust transformation and variable selection: A unified approach. *Canad. J. Statist.* **37** 361–380. MR2547204 https://doi.org/10.1002/cjs.10021

HAMPEL, F. R. (1975). Beyond location parameters: Robust concepts and methods. *Bull. Int. Stat. Inst.* **46** 375–382. MR0483172 https://doi.org/10.1016/0370-2693(74)90750-3

HARRELL, F. E. JR. (2019). Hmisc. R package version 4.2-0.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HINKLEY, D. V. and RUNGER, G. (1984). The analysis of transformed data. *J. Amer. Statist. Assoc.* **79** 302–320. MR0755087

HOYLE, M. H. (1973). Transformations—An introduction and a bibliography. *Int. Stat. Rev.* **41** 203–223. MR0423611 https://doi.org/10.2307/1402836

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

JOHN, J. A. and DRAPER, N. R. (1980). An alternative family of transformations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **29** 190–197.

KLEINBAUM, D. G. and KUPPER, L. (1978). *Applied Regression Analysis and Other Multivariable Methods*. Duxbury, Boston, MA.

LAWRANCE, A. J. (1987). The score statistic for regression transformation. *Biometrika* **74** 275–279. MR0903128 https://doi.org/10.1093/biomet/74.2.275

MARAZZI, A., VILLAR, A. J. and YOHAI, V. J. (2009). Robust response transformations based on optimal prediction. *J. Amer. Statist. Assoc.* **104** 360–370. MR2663044 https://doi.org/10.1198/jasa.2009.0109

MCCULLAGH, P. (2002a). Comment on "Box–Cox transformations in linear models: Large sample theory and tests of normality" by Chen, Lockhart and Stephens. *Canad. J. Statist.* **30** 212–213.

MCCULLAGH, P. (2002b). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320 https://doi.org/10.1214/aos/1035844977

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

NYCHKA, D. and RUPPERT, D. (1995). Nonparametric transformations for both sides of a regression model. *J. Roy. Statist. Soc. Ser. B* **57** 519–532. MR1341322

PERICCHI, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika* **68** 35–43. MR0614939 https://doi.org/10.1093/biomet/68.1.35

PERROTTA, D., RIANI, M. and TORTI, F. (2009). New robust dynamic plots for regression mixture detection. *Adv. Data Anal. Classif.* **3** 263–279. MR2570505 https://doi.org/10.1007/s11634-009-0050-y

PREGIBON, D. (1980). Goodness of link tests for generalized linear models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **29** 15–23, 14.

PROIETTI, T. and RIANI, M. (2009). Transformations and seasonal adjustment. *J. Time Series Anal.* **30** 47–69. MR2488635 https://doi.org/10.1111/j.1467-9892.2008.00600.x

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3** 425–461.

REID, N. (2002). Comment on "Box–Cox transformations in linear models: Large sample theory and tests of normality" by Chen, Lockhart and Stephens. *Canad. J. Statist.* **30** 211.

RIANI, M. (2009). Robust transformations in univariate and multivariate time series. *Econometric Rev.* **28** 262–278. MR2655628 https://doi.org/10.1080/07474930802388074

RIANI, M., ATKINSON, A. C. and CERIOLI, A. (2009). Finding an unknown number of multivariate outliers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 447–466. MR2649609 https://doi.org/10.1111/j.1467-9868.2008.00692.x

RIANI, M., CERIOLI, A., ATKINSON, A. C. and PERROTTA, D. (2014). Monitoring robust regression. *Electron. J. Stat.* **8** 646–677. MR3211027 https://doi.org/10.1214/14-EJS897

ROSS, G. J. S. (1990). *Nonlinear Estimation*. Springer, New York.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

ROYSTON, P. J. and ALTMAN, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *J. R. Stat. Soc. Ser. C. Appl. Stat.* **43** 429–467.

SAKIA, R. M. (1992). The Box–Cox transformation technique: A review. *Statistician* **41** 169–178.

SEO, H. S. (2019). A robust method for response variable transformations using dynamic plots. *Commun. Stat. Appl. Methods* **26** 463–471. https://doi.org/10.29220/CSAM.2019.26.5.463

SONG, X.-Y. and LU, Z.-H. (2012). Semiparametric transformation models with Bayesian P-splines. *Stat. Comput.* **22** 1085–1098. MR2950087 https://doi.org/10.1007/s11222-011-9280-x

SPECTOR, P., FRIEDMAN, J., TIBSHIRANI, R., LUMLEY, T., GARBETT, S. and BARON, J. (2016). acepack: ACE and AVAS for selecting multiple regression transformations. R package version 1.4.1.

SWEETING, T. J. (1984). On the choice of prior distribution for the Box–Cox transformed linear model. *Biometrika* **71** 127–134. MR0738333 https://doi.org/10.1093/biomet/71.1.127

TAYLOR, J. M. G. (1986). The retransformed mean after a fitted power transformation. *J. Amer. Statist. Assoc.* **81** 114–118.

TAYLOR, J. M. G. (2004). Transformations II. *Wiley Statsref.*

TAYLOR, J. M. G. and LIU, N. (2007). Statistical issues involved with extending standard models. In *Advances in Statistical Modeling and Inference*. *Ser. Biostat.* **3** 299–311. World Sci. Publ., Hackensack, NJ. MR2416121 https://doi.org/10.1142/9789812708298_0015

TIBSHIRANI, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *J. Amer. Statist. Assoc.* **83** 394–405. MR0971365

VELILLA, S. (1995). Diagnostics and robust estimation in multivariate data transformations. *J. Amer. Statist. Assoc.* **90** 945–951. MR1354010

WANG, D. and MURPHY, M. (2005). Identifying nonlinear relationships in regression using the ACE algorithm. *J. Appl. Stat.* **32** 243–258. MR2140333 https://doi.org/10.1080/02664760500054517

WANG, N. and RUPPERT, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *J. Amer. Statist. Assoc.* **90** 522–534. MR1340505

WEISBERG, S. (2005). Yeo–Johnson power transformations. Available at https://www.stat.umn.edu/arc/yjpower.pdf.

WIENS, B. L. (1999). When log-normal and gamma models give different results: A case study. *Amer. Statist.* **53** 89–93.

YANG, Z. (2006). A modified family of power transformations. *Econom. Lett.* **92** 14–19. MR2237746 https://doi.org/10.1016/j.econlet.2006.01.011

YEO, I.-K. and JOHNSON, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87** 954–959. MR1813988 https://doi.org/10.1093/biomet/87.4.954

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431 https://doi.org/10.1214/aos/1176350366

ZHANG, T. and YANG, B. (2017). Box–Cox transformation in big data. *Technometrics* **59** 189–201. MR3635042 https://doi.org/10.1080/00401706.2016.1156025

Statistical Science
2021, Vol. 36, No. 2, 256–263
https://doi.org/10.1214/20-STS780

Noncommutative Probability and Multiplicative Cascades

Ian W. McKeague

Abstract. Various aspects of standard model particle physics might be explained by a suitably rich algebra acting on itself, as suggested by Furey (2015). The present paper develops the asymptotics of large causal tree diagrams that combine freely independent elements in such an algebra. The Marčenko–Pastur law and Wigner's semicircle law are shown to emerge as limits of normalized sum-over-paths of nonnegative elements assigned to the edges of causal trees. These results are established in the setting of noncommutative probability. Trees with classically independent positive edge weights (random multiplicative cascades) were originally proposed by Mandelbrot as a model displaying the fractal features of turbulence. The novelty of the present work is the use of noncommutative (free) probability to allow the edge weights to take values in an algebra. An application to theoretical neuroscience is also discussed.

Key words and phrases: Mandelbrot cascade, Marčenko–Pastur law, martingale convergence, random matrices, Wigner's semicircle law.

REFERENCES

ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2009). *An Introduction to Random Matrices*, 1st ed. *Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

ASCOLI, G. A. (2015). *Trees of the Brain*, *Roots of the Mind*, 1st ed. MIT Press, Cambridge.

ATTAL, S. and NECHITA, I. (2011). Discrete approximation of the free Fock space. In *Séminaire de Probabilités XLIII*. *Lecture Notes in Math.* **2006** 379–394. Springer, Berlin. MR2790382 https://doi.org/10.1007/978-3-642-15217-7_16

BARRAL, J. (2014). Mandelbrot cascades and related topics. In *Geometry and Analysis of Fractals*. *Springer Proc. Math. Stat.* **88** 1–45. Springer, Heidelberg. MR3275997 https://doi.org/10.1007/978-3-662-43920-3_1

BARRAL, J., KUPIAINEN, A., NIKULA, M., SAKSMAN, E. and WEBB, C. (2014). Critical Mandelbrot cascades. *Comm. Math. Phys.* **325** 685–711. MR3148099 https://doi.org/10.1007/s00220-013-1829-4

BERCOVICI, H. and VOICULESCU, D. (1993). Free convolution of measures with unbounded support. *Indiana Univ. Math. J.* **42** 733–773. MR1254116 https://doi.org/10.1512/iumj.1993.42.42033

BOMBELLI, L., LEE, J., MEYER, D. and SORKIN, R. D. (1987). Space-time as a causal set. *Phys. Rev. Lett.* **59** 521–524. MR0899046 https://doi.org/10.1103/PhysRevLett.59.521

BRAGA, G. A., SANCHIS, R. and SCHIEBER, T. A. (2005). Critical percolation on a Bethe lattice revisited. *SIAM Rev.* **47** 349–365. MR2179901 https://doi.org/10.1137/S0036144503433685

BROADBENT, S. R. and HAMMERSLEY, J. M. (1957). Percolation processes. I. Crystals and mazes. *Proc. Camb. Philos. Soc.* **53** 629–641. MR0091567 https://doi.org/10.1017/s0305004100032680

BROCHINI, L., DE ANDRADE COSTA, A., ABADI, M., ROQUE, A. C., STOL, J. and KINOUCHI, O. (2016). Phase transitions and self-organized criticality in networks of stochastic spiking neurons. *Sci. Rep.* **6** 35831.

CHISTYAKOV, G. P. and GÖTZE, F. (2008). Limit theorems in free probability theory. I. *Ann. Probab.* **36** 54–90. MR2370598 https://doi.org/10.1214/009117907000000051

COLLINS, B., HAYDEN, P. and NECHITA, I. (2017). Random and free positive maps with applications to entanglement detection. *Int. Math. Res. Not. IMRN* **2017** 869–894. MR3658153 https://doi.org/10.1093/imrn/rnw054

CORTÊS, M. and SMOLIN, L. (2014a). Quantum energetic causal sets. *Phys. Rev. D* **90** 044035.

CORTÊS, M. and SMOLIN, L. (2014b). The universe as a process of unique events. *Phys. Rev. D* **90** 084007.

DERRIDA, B. and SPOHN, H. (1988). Polymers on disordered trees, spin glasses, and traveling waves **51** 817–840. MR0971033 https://doi.org/10.1007/BF01014886

DOWKER, F. (2005). Causal sets and the deep structure of spacetime. In 100 *Years of Relativity* 445–464. World Sci. Publ., Hackensack, NJ. MR2223124 https://doi.org/10.1142/9789812700988_0016

DURRETT, R. and LIGGETT, T. M. (1983). Fixed points of the smoothing transformation. *Z. Wahrsch. Verw. Gebiete* **64** 275–301. MR0716487 https://doi.org/10.1007/BF00532962

FISHER, M. P. A. (2015). Quantum cognition: The possibility of processing with nuclear spins in the brain. *Ann. Physics* **362** 593–602. MR3411643 https://doi.org/10.1016/j.aop.2015.08.020

FUREY, C. (2015). Standard model physics from an algebra? Ph.D. thesis, University of Waterloo. Available at arXiv:1611.09182 [hep-th].

Ian W. McKeague is Professor of Biostatistics, Department of Biostatistics, Columbia University, New York, New York 10032, USA (e-mail: ian.mckeague@columbia.edu).

HUANG, C. (2016). Moments for multidimensional Mandelbrot cascades. *J. Appl. Probab.* **53** 1–21. MR3540776 https://doi.org/10.1017/jpr.2015.4

KAHANE, J.-P. and PEYRIÈRE, J. (1976). Sur certaines martingales de Benoit Mandelbrot. *Adv. Math.* **22** 131–145. MR0431355 https://doi.org/10.1016/0001-8708(76)90151-1

KARGIN, V. (2013). Lecture notes on free probability. Available at arXiv:1305.2611 [math.PR].

LIU, Q. and ROUAULT, A. (2000). Limit theorems for Mandelbrot's multiplicative cascades. *Ann. Appl. Probab.* **10** 218–239. MR1765209 https://doi.org/10.1214/aoap/1019737670

MANDELBROT, B. (1974a). Multiplications aléatoires itérées et distributions invariantes par moyenne pondérée aléatoire. *C. R. Acad. Sci. Paris, Sér. A* **278** 289–292. MR0431351

MANDELBROT, B. B. (1974b). Intermittent turbulence in self-similar cascades, divergence of high moments and dimension of the carrier. *J. Fluid Mech.* **62** 331–358.

MARKOPOULOU, F. (2000). The internal description of a causal set: What the universe looks like from the inside. *Comm. Math. Phys.* **211** 559–583. MR1773808 https://doi.org/10.1007/s002200050826

MINGO, J. A. and SPEICHER, R. (2017). *Free Probability and Random Matrices*. *Fields Institute Monographs* **35**. Springer, New York. MR3585560 https://doi.org/10.1007/978-1-4939-6942-5

NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. *London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879 https://doi.org/10.1017/CBO9780511735127

OSSIANDER, M. and WAYMIRE, E. C. (2002). On estimation theory for multiplicative cascades. *Sankhyā Ser. A* **64** 323–343. MR1981762

PISIER, G. (2016). *Martingales in Banach Spaces*, 1st ed. *Cambridge Studies in Advanced Mathematics* **155**. Cambridge Univ. Press, Cambridge. MR3617459

ROSENBLATT, M. and VAN ATTA, C., eds. (1972). Statistical models and turbulence. In *Proceedings of a Symposium Held at the University of California, San Diego (La Jolla), Calif., July* 15–21, 1971. *Lecture Notes in Physics* **12**. Springer, Berlin. MR0438885

SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. Springer, New York. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

TAO, T. (2010). Free probability. Available at terrytao.wordpress.com, 254A, Notes 5.

TURING, A. M. (1950). Computing machinery and intelligence. *Mind* **59** 433–460. MR0037064 https://doi.org/10.1093/mind/LIX.236.433

Statistical Science
2021, Vol. 36, No. 2, 264–290
https://doi.org/10.1214/20-STS783

A Selective Overview of Deep Learning

Jianqing Fan, Cong Ma and Yiqiao Zhong

Abstract. Deep learning has achieved tremendous success in recent years. In simple words, deep learning uses the composition of many nonlinear functions to model the complex dependency between input features and labels. While neural networks have a long history, recent advances have significantly improved their empirical performance in computer vision, natural language processing and other predictive tasks. From the statistical and scientific perspective, it is natural to ask: What is deep learning? What are the new characteristics of deep learning, compared with classical statistical methods? What are the theoretical foundations of deep learning?

To answer these questions, we introduce common neural network models (e.g., convolutional neural nets, recurrent neural nets, generative adversarial nets) and training techniques (e.g., stochastic gradient descent, dropout, batch normalization) from a statistical point of view. Along the way, we highlight new characteristics of deep learning (including depth and overparametrization) and explain their practical and theoretical benefits. We also sample recent results on theories of deep learning, many of which are only suggestive. While a complete understanding of deep learning remains elusive, we hope that our perspectives and discussions serve as a stimulus for new statistical research.

Key words and phrases: Neural networks, overparametrization, stochastic gradient descent, approximation theory, generalization error.

REFERENCES

ABADI, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

ABBASI-ASL, R., CHEN, Y., BLONIARZ, A., OLIVER, M., WILLMORE, B. D., GALLANT, J. L. and YU, B. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *BioRxiv* 465534.

ALLEN-ZHU, Z. and LI, Y. (2019). Can SGD learn recurrent neural networks with provable generalization? Preprint. Available at arXiv:1902.01028.

ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2018). A convergence theory for deep learning via over-parameterization. Preprint. Available at arXiv:1811.03962.

ANTHONY, M. and BARTLETT, P. L. (2009). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038 https://doi.org/10.1017/CBO9780511624216

ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the* 34*th International Conference on Machine Learning*. *PMLR* **70** 214–223.

ARNOLD, V. I. (2009). On functions of three variables. In *Collected Works*: *Representations of Functions*, *Celestial Mechanics and KAM Theory*, 1957–1965 5–8.

ARORA, S., GE, R., LIANG, Y., MA, T. and ZHANG, Y. (2017). Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 224–232. JMLR.org.

ARORA, S., DU, S. S., HU, W., LI, Z. and WANG, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. Preprint. Available at arXiv:1901.08584.

BAI, Y., MA, T. and RISTESKI, A. (2018). Approximability of discriminators implies diversity in GANs. Preprint. Available at arXiv:1806.10586.

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans*. *Inf*. *Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500

BARTLETT, P. L. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans*. *Inf*. *Theory* **44** 525–536. MR1607706 https://doi.org/10.1109/18.661502

BARTLETT, P. L., FOSTER, D. J. and TELGARSKY, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems* **30** (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 6240–6249. Curran Associates, Red Hook.

Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Department of ORFE, Princeton University, Princeton, New Jersey, 08544, USA (e-mail: jqfan@princeton.edu). Cong Ma is PhD, Department of ORFE, Princeton University, Princeton, New Jersey, 08544, USA (e-mail: congm@princeton.edu). Yiqiao Zhong is postdoc, Electrical Engineering Department, Stanford University, Stanford, California, 94305, USA (e-mail: yiqiaoz@stanford.edu).

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. MR3953451 https://doi.org/10.1214/18-AOS1747

BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

BOTTOU, L., ed. (1998) Online learning and stochastic approximations. In *Online Learning in Neural Networks. Publications of the Newton Institute* **17**. Cambridge Univ. Press, Cambridge.

BOUSQUET, O. and ELISSEEFF, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526. MR1929416 https://doi.org/10.1162/153244302760200704

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

BREIMAN, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957 https://doi.org/10.1214/aos/1032181158

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

CAO, C., LIU, F., TAN, H., SONG, D., SHU, W., LI, W., ZHOU, Y., BO, X. and XIE, Z. (2018). Deep learning and its applications in biomedicine. *Genomics Proteomics Bioinform.* **16** 17–32.

CHEN, S., DOBRIBAN, E. and LEE, J. H. (2019). Invariance reduces variance: Understanding data augmentation in deep learning and beyond. Preprint. Available at arXiv:1907.10905.

CHEN, T., RUBANOVA, Y., BETTENCOURT, J. and DUVENAUD, D. (2018). Neural ordinary differential equations. Preprint. Available at arXiv:1806.07366.

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2019a). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Program.* **176** 5–37. MR3960803 https://doi.org/10.1007/s10107-019-01363-6

CHEN, Y., CHI, Y., FAN, J., MA, C. and YAN, Y. (2019b). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. Preprint. Available at arXiv:1902.07698.

CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* 3040–3050.

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. and BENGIO, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Preprint. Available at arXiv:1406.1078.

CORTES, C., GONZALVO, X., KUZNETSOV, V., MOHRI, M. and YANG, S. (2017). Adanet: Adaptive structural learning of artificial neural networks. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 874–883. JMLR.org.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

DEVROYE, L. P. and WAGNER, T. J. (1979). Distribution-free performance bounds for potential function rules. *IEEE Trans. Inf. Theory* **25** 601–604. MR0545015 https://doi.org/10.1109/TIT.1979.1056087

DE FAUW, J., LEDSAM, J. R., ROMERA-PAREDES, B., NIKOLOV, S., TOMASEV, N., BLACKWELL, S., ASKHAM, H., GLOROT, X., O'DONOGHUE, B. et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24** 1342.

DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math. Chall. Lect.* **1** 32.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

DU, S. S., LEE, J. D., LI, H., WANG, L. and ZHAI, X. (2018). Gradient descent finds global minima of deep neural networks. Preprint. Available at arXiv:1811.03804.

DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. MR2825422

E, W., HAN, J. and JENTZEN, A. (2017). Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.* **5** 349–380. MR3736669 https://doi.org/10.1007/s40304-017-0117-6

E, W., MA, C. and WANG, Q. (2019). A priori estimates of the population risk for residual networks. Preprint. Available at arXiv:1903.02154.

ELDAN, R. and SHAMIR, O. (2016). The power of depth for feedforward neural networks. In *Conference on Learning Theory* 907–940.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FELDMAN, V. and VONDRAK, J. (2019). High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. Preprint. Available arXiv:1902.10710.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

FU, H., CHI, Y. and LIANG, Y. (2018). Local geometry of one-hidden-layer neural networks for logistic regression. Preprint. Available at arXiv:1802.06463.

FUKUSHIMA, K. and MIYAKE, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets* 267–285. Springer, Berlin.

GAO, C., LIU, J., YAO, Y. and ZHU, W. (2018). Robust estimation and generative adversarial nets. Preprint. Available at arXiv:1810.02030.

GOEL, S., KLIVANS, A. and MEKA, R. (2018). Learning one convolutional layer with overlapping patches. Preprint. Available at arXiv:1802.02547.

GOLOWICH, N., RAKHLIN, A. and SHAMIR, O. (2020). Size-independent sample complexity of neural networks. *Inf. Inference* **9** 473–504. MR4108976 https://doi.org/10.1093/imaiai/iaz007

GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*, 4th ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR3024913

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.

GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018a). Characterizing implicit bias in terms of optimization geometry. Preprint. Available at arXiv:1802.08246.

GUNASEKAR, S., LEE, J. D., SOUDRY, D. and SREBRO, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems* 9482–9491.

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 https://doi.org/10.1214/aos/1176349020

HARDT, M., RECHT, B. and SINGER, Y. (2015). Train faster, generalize better: Stability of stochastic gradient descent. Preprint. Available at rXiv:1509.01240.

HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. Preprint. Available at arXiv:1903.08560.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016b). Identity mappings in deep residual networks. In *European Conference on Computer Vision* 630–645. Springer, Berlin.

HINTON, G., SRIVASTAVA, N. and SWERSKY, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. Preprint. Available at arXiv:1207.0580.

HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Comput*. **9** 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw*. **4** 251–257.

HUANG, G., LIU, Z., VAN DER MAATEN, L. and WEINBERGER, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708.

HUBEL, D. H. and WIESEL, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol*. **160** 106–154.

IANDOLA, F. N., HAN, S., MOSKEWICZ, M. W., ASHRAF, K., DALLY, W. J. and KEUTZER, K. (2016). SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. Preprint. Available at arXiv:1602.07360.

IOFFE, S. and SZEGEDY, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Preprint. Available at arXiv:1502.03167.

JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems* 8580–8589.

JAIN, P., KAKADE, S. M., KIDAMBI, R., NETRAPALLI, P. and SIDFORD, A. (2017). Accelerating stochastic gradient descent. Preprint. Available at arXiv:1704.08227.

JAVANMARD, A., MONDELLI, M. and MONTANARI, A. (2019). Analysis of a two-layer neural network via displacement convexity. Preprint. Available at arXiv:1901.01375.

JI, Z. and TELGARSKY, M. (2018). Risk and parameter convergence of logistic regression. Preprint. Available at arXiv:1803.07300.

KIDAMBI, R., NETRAPALLI, P., JAIN, P. and KAKADE, S. (2018). On the insufficiency of existing momentum schemes for stochastic optimization. In 2018 *Information Theory and Applications Workshop* (*ITA*) 1–9. IEEE.

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat*. **23** 462–466. MR0050243 https://doi.org/10.1214/aoms/1177729392

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational bayes. Preprint. Available at arXiv:1312.6114.

KLUSOWSKI, J. M. and BARRON, A. R. (2016). Risk bounds for high-dimensional ridge function combinations including neural networks. Preprint. Available at arXiv:1607.01434.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.

KUSHNER, H. J. and YIN, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*: *Stochastic Modelling and Applied Probability*, 2nd ed. *Applications of Mathematics* (*New York*) **35**. Springer, New York. MR1993642

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444. https://doi.org/10.1038/nature14539

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.

LI, Y., SWERSKY, K. and ZEMEL, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning* 1718–1727.

LI, H., XU, Z., TAYLOR, G., STUDER, C. and GOLDSTEIN, T. (2018a). Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems* 6391–6401.

LI, X., LU, J., WANG, Z., HAUPT, J. and ZHAO, T. (2018b). On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond. Preprint. Available at arXiv:1806.05159.

LIANG, T. (2017). How well can generative adversarial networks (GAN) learn densities: A nonparametric view. Preprint. Available at arXiv:1712.08244.

LIN, M., CHEN, Q. and YAN, S. (2013). Network in network. Preprint. Available at arXiv:1312.4400.

LIN, H. W., TEGMARK, M. and ROLNICK, D. (2017). Why does deep and cheap learning work so well? *J. Stat. Phys*. **168** 1223–1247. MR3691248 https://doi.org/10.1007/s10955-017-1836-5

MAAS, A. L., HANNUN, A. Y. and NG, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML* **30** 3.

MAIOROV, V. E. and MEIR, R. (2000). On the near optimality of the stochastic approximation of smooth functions by neural networks. *Adv. Comput. Math*. **13** 79–103. MR1759189 https://doi.org/10.1023/A:1018993908478

MAKOVOZ, Y. (1996). Random approximants and neural networks. *J. Approx. Theory* **85** 98–109. MR1382053 https://doi.org/10.1006/jath.1996.0031

MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. Preprint. Available at arXiv:1902.06015.

MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115

MHASKAR, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput*. **8** 164–177.

MHASKAR, H., LIAO, Q. and POGGIO, T. (2016). Learning functions: When is deep better than shallow. Preprint. Available at arXiv:1603.00988.

MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529.

MONDELLI, M. and MONTANARI, A. (2018). On the connection between learning two-layers neural networks and tensor decomposition. Preprint. Available at arXiv:1802.07301.

NESTEROV, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR* **269** 543–547. MR0701288

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015). Norm-based capacity control in neural networks. In *Conference on Learning Theory* 1376–1401.

NOWOZIN, S., CSEKE, B. and TOMIOKA, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems* 271–279.

PARBERRY, I. (1994). *Circuit Complexity and Neural Networks*. *Foundations of Computing Series*. MIT Press, Cambridge, MA. MR1286941

PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., DEVITO, Z., LIN, Z., DESMAISON, A., ANTIGA, L. et al. (2017). Automatic differentiation in PyTorch.

PINKUS, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numer.* **8** 143–195. MR1819645 https://doi.org/10.1017/S0962492900002919

POGGIO, T., MHASKAR, H., ROSASCO, L., MIRANDA, B. and LIAO, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *Int. J. Autom. Comput.* **14** 503–519.

POLJAK, B. T. (1964). Some methods of speeding up the convergence of iterative methods. *USSR Comput. Math. Math. Phys.* **4** 1–17.

POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046

POLYAK, B. T. and TSYPKIN, Y. Z. (1979). Adaptive estimation algorithms (convergence, optimality, stability). *Autom. Remote Control* **3** 71–84. MR0544876

POULTNEY, C., CHOPRA, S., LECUN, Y. et al. (2007). Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems* 1137–1144.

REDDI, S. J., KALE, S. and KUMAR, S. (2018). On the convergence of adam and beyond.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

ROGERS, W. H. and WAGNER, T. J. (1978). A finite sample distribution-free performance bound for local discrimination rules. *Ann. Statist.* **6** 506–514. MR0468042

ROLNICK, D. and TEGMARK, M. (2017). The power of deeper networks for expressing natural functions. Preprint. Available at arXiv:1705.05502.

ROMANO, Y., SESIA, M. and CANDÈS, E. J. (2018). Deep knockoffs. Preprint. Available at arXiv:1811.06687.

ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. Preprint. Available at arXiv:1805.00915.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ. San Diego La Jolla Inst. for Cognitive Science.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A. et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** 211–252. https://doi.org/10.1007/s11263-015-0816-y

SAK, H., SENIOR, A. and BEAUFAYS, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

SALAKHUTDINOV, R. and HINTON, G. (2009). Deep Boltzmann machines. In *Artificial Intelligence and Statistics* 448–455.

SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A. and CHEN, X. (2016). Improved techniques for training GANs. In *Advances in Neural Information Processing Systems* 2234–2242.

SCHMIDT-HIEBER, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. Preprint. Available at arXiv:1708.06633.

SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning*: *From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

SHALEV-SHWARTZ, S., SHAMIR, O., SREBRO, N. and SRIDHARAN, K. (2010). Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* **11** 2635–2670. MR2738779

SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M. et al. (2017). Mastering the game of go without human knowledge. *Nature* **550** 354.

SILVERMAN, B. W. (1998). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

SINGH, C., MURDOCH, W. J. and YU, B. (2018). Hierarchical interpretations for neural network predictions. Preprint. Available at arXiv:1806.05337.

SIRIGNANO, J. and SPILIOPOULOS, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.* **80** 725–752. MR4074020 https://doi.org/10.1137/18M1192184

SOLTANOLKOTABI, M. (2017). Learning relus via gradient descent. In *Advances in Neural Information Processing Systems* 2007–2017.

SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Paper No. 70, 57. MR3899772

SPRECHER, D. A. (1965). On the structure of continuous functions of several variables. *Trans. Amer. Math. Soc.* **115** 340–355. MR0210852 https://doi.org/10.2307/1994273

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566 https://doi.org/10.1214/aos/1176349548

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184. MR1272079 https://doi.org/10.1214/aos/1176325361

SUTSKEVER, I., MARTENS, J., DAHL, G. and HINTON, G. (2013). On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning* 1139–1147.

SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I. and FERGUS, R. (2013). Intriguing properties of neural networks. Preprint. Available at arXiv:1312.6199.

SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V. and RABINOVICH, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1–9.

TELGARSKY, M. (2016). Benefits of depth in neural networks. Preprint. Available at arXiv:1602.04485.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAPNIK, V. N. and CHERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VINCENT, P., LAROCHELLE, H., BENGIO, Y. and MANZAGOL, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* 1096–1103. ACM, New York.

WAGER, S., WANG, S. and LIANG, P. S. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems* 351–359.

WILSON, A. C., ROELOFS, R., STERN, M., SREBRO, N. and RECHT, B. (2017). The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems* **30** (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 4148–4158. Curran Associates, Red Hook.

WU, Y., SCHUSTER, M., CHEN, Z., LE, Q. V., NOROUZI, M., MACHEREY, W., KRIKUN, M., CAO, Y., GAO, Q. et al. (2016).

Google's neural machine translation system: Bridging the gap between human and machine translation. Preprint. Available at arXiv:1609.08144.

YOSINSKI, J., CLUNE, J., BENGIO, Y. and LIPSON, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328.

YOSINSKI, J., CLUNE, J., NGUYEN, A., FUCHS, T. and LIPSON, H. (2015). Understanding neural networks through deep visualization. Preprint. Available at arXiv:1506.06579.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. Preprint. Available at arXiv:1611.03530.

ZHONG, K., SONG, Z., JAIN, P., BARTLETT, P. L. and DHILLON, I. S. (2017). Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 4140–4149. JMLR.org.

Statistical Science
2021, Vol. 36, No. 2, 291–302
https://doi.org/10.1214/20-STS784

Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications

Tze Leung Lai and Hongsong Yuan

Abstract. Stochastic approximation was introduced in 1951 to provide a new theoretical framework for root finding and optimization of a regression function in the then-nascent field of statistics. This review shows how it has evolved in response to other developments in statistics, notably time series and sequential analysis, and to applications in artificial intelligence, economics and engineering. Its resurgence in the big data era has led to new advances in both theory and applications of this microcosm of statistics and data science.

Key words and phrases: Control, gradient boosting, optimization, recursive stochastic algorithms, regret, weak greedy variable selection.

REFERENCES

ANDERSON, T. W. and TAYLOR, J. B. (1976). Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica* **44** 1289–1302.

ÅSTÖM, K. J. and WITTENMARK, B. (1973). On self-tuning regulators. *Automatica* **9** 189–199.

ÅSTRÖM, K. J. and WITTENMARK, B. (1971). Problems of identification and control. *J. Math. Anal. Appl.* **34** 90–113. MR0278799 https://doi.org/10.1016/0022-247X(71)90161-2

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BLUM, J. R. (1954). Approximation methods which converge with probability one. *Ann. Math. Stat.* **25** 382–386. MR0062399 https://doi.org/10.1214/aoms/1177728794

BONYADI, M. and MICHALEWICZ, Z. (2016). Analysis of stability, local convergence, and transformation sensitivity of a variant of particle swarm optimization algorithm. *IEEE Trans. Evol. Comput.* **20** 370–385.

BOX, G. E. P. and WILSON, K. B. (1951). On the experimental attainment of optimum conditions. *J. Roy. Statist. Soc. Ser. B* **13** 1–38; discussion: 38–45. MR0046009

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

CHOI, K. P., LAI, T. L., TONG, X. T. and WONG, W. K. (2021). A statistical approach to adaptive parameter tuning in nature-inspired optimization and optimal sequential design of dose-finding trials. *Statist. Sinica* **31**. To appear.

CLERC, M. (2006). *Particle Swarm Optimization*. ISTE, London. MR2269598 https://doi.org/10.1002/9780470612163

DE LA PEÑA, V. H., KLASS, M. J. and LAI, T. L. (2009). Theory and applications of multivariate self-normalized processes. *Stochastic Process. Appl.* **119** 4210–4227. MR2565565 https://doi.org/10.1016/j.spa.2009.10.003

EFRON, B. (2003). Robbins, empirical Bayes and microarrays. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 366–378. MR1983533 https://doi.org/10.1214/aos/1051027871

FREUND, Y. and SCHAPIRE, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** 337–407. With discussion and a rejoinder by the authors. MR1790002 https://doi.org/10.1214/aos/1016218223

FUCHS, J.-J. J. (1982). Indirect stochastic adaptive control: The general delay-white noise case. *IEEE Trans. Automat. Control* **27** 219–223. MR0673095 https://doi.org/10.1109/TAC.1982.1102827

GOODWIN, G. C., RAMADGE, P. J. and CAINES, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19** 829–853. MR0634955 https://doi.org/10.1137/0319052

GUO, L., HUANG, D. W. and HANNAN, E. J. (1990). On ARX(∞) approximation. *J. Multivariate Anal.* **32** 17–47. MR1035605 https://doi.org/10.1016/0047-259X(90)90069-T

HANNAN, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2** 135–161. With comments and a reply by the author. MR0904031

HASSABIS, D., KUMARAN, D., SUMMERFIELD, C. and BOTVINICK, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* **95** 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

HOTELLING, H. (1941). Experimental determination of the maximum of a function. *Ann. Math. Stat.* **12** 20–45. MR0003521 https://doi.org/10.1214/aoms/1177731784

Tze Leung Lai is Professor, Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, California 94305, USA (e-mail: lait@stanford.edu). Hongsong Yuan is Associate Professor, School of Information Management and Engineering, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China (e-mail: yuan.hongsong@shufe.edu.cn).

ING, C.-K. and LAI, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* **21** 1473–1513. MR2895106 https://doi.org/10.5705/ss.2010.081

KENNEDY, J. and EBERHART, R. (1995). Particle swarm optimization. In *Proc. IEEE Int. Conf. Neural Networks* 4 1942–1048, Perth, Australia.

KIEFER, J. (1953). Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.* **4** 502–506. MR0055639 https://doi.org/10.2307/2032161

KIEFER, J. (1957). Optimum sequential search and approximation methods under minimum regularity assumptions. *J. Soc. Indust. Appl. Math.* **5** 105–136. MR0092326

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23** 462–466. MR0050243 https://doi.org/10.1214/aoms/1177729392

KUMAR, P. R. and VARAIYA, P. (1986). *Stochastic Systems*, 1st ed. ed. Prentice-Hall, Englewood Cliffs, NJ. MR3471643

KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Applied Mathematical Sciences* **26**. Springer, New York-Berlin. MR0499560

LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114. MR0902248 https://doi.org/10.1214/aos/1176350495

LAI, T. L. (1989). Extended stochastic Lyapunov functions and recursive algorithms in linear stochastic systems. In *Stochastic Differential Systems* (*Bad Honnef*, 1988) (N. Christopeit, K. Helmes and N. Kohlmann, eds.). *Lect. Notes Control Inf. Sci.* **126** 206–220. Springer, Berlin. MR1236069 https://doi.org/10.1007/BFb0043786

LAI, T. L. (2003). Stochastic approximation. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 391–406. MR1983535 https://doi.org/10.1214/aos/1051027873

LAI, T. L., CHOI, A. and TSANG, K. W. (2019). Statistical science in information technology and precision medicine. *Ann. Math. Sci. Appl.* **4** 413–438. MR4020370 https://doi.org/10.4310/AMSA.2019.v4.n2.a6

LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196–1221. MR0550144

LAI, T. L. and ROBBINS, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Z. Wahrsch. Verw. Gebiete* **56** 329–360. MR0621117 https://doi.org/10.1007/BF00536178

LAI, T. L. and ROBBINS, H. (1982a). Adaptive design and the multiperiod control problem. In *Statistical Decision Theory and Related Topics, III, Vol.* 2 (*West Lafayette, Ind.*, 1981) 103–120. Academic Press, New York. MR0705310

LAI, T. L. and ROBBINS, H. (1982b). Iterated least squares in multiperiod control. *Adv. in Appl. Math.* **3** 50–73. MR0646499 https://doi.org/10.1016/S0196-8858(82)80005-5

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

LAI, T. L. and SIEGMUND, D. (1986). The contributions of Herbert Robbins to mathematical statistics. *Statist. Sci.* **1** 276–284. MR0846004

LAI, T. L. and WEI, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** 154–166. MR0642726

LAI, T. L. and WEI, C. Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Trans. Automat. Control* **31** 898–906. MR0855543 https://doi.org/10.1109/TAC.1986.1104138

LAI, T. L., XIA, T. and YUAN, H. (2020). Modified gradient boosting in high-dimensional nonlinear regression. Technical report, Dept. Statistics, Stanford Univ.

LAI, T. L., XU, H. and YUAN, H. (2020). Self-normalized martingales, recursive orthogonal matching pursuit and modified gradient boosting in high-dimensional stochastic regression models. Technical report, Dept. Statistics, Stanford Univ.

LAI, T. L. and YING, Z. (1991a). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Control Optim.* **29** 1061–1090. MR1110087 https://doi.org/10.1137/0329058

LAI, T. L. and YING, Z. (1991b). Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM J. Control Optim.* **29** 1091–1127. MR1110088 https://doi.org/10.1137/0329059

LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **AC-22** 551–575. MR0465458 https://doi.org/10.1109/tac.1977.1101561

LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919

MALLAT, S. and ZHANG, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41** 3397–3415.

NEMIROVSKI, A. and YUDIN, D. (1978). On Cezari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Sov. Math., Dokl.* **19**. MR0482494

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *A Wiley-Interscience Publication*. Wiley, New York. MR0702836

PAGE, W. (1984). An interview with Herbert Robbins. *College Math. J.* **15** 2–24. MR0759951 https://doi.org/10.2307/3027425

PELIGRAD, M., SANG, H., ZHONG, Y. and WU, W. B. (2014). Exact moderate and large deviations for linear processes. *Statist. Sinica* **24** 957–969. MR3235407

POLYAK, B. T. (1990). A new method of stochastic approximation type. *Avtomat. i Telemekh.* **7** 98–107. MR1071220

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766 https://doi.org/10.1214/11-EJS631

ROBBINS, H. E. (1944). Two properties of the function $\cos x$. *Bull. Amer. Math. Soc.* **50** 750–752. MR0011342 https://doi.org/10.1090/S0002-9904-1944-08232-3

ROBBINS, H. E. (1945). On the measure of a random set. II. *Ann. Math. Stat.* **16** 342–347. MR0019239 https://doi.org/10.1214/aoms/1177731060

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (*Proc. Sympos., Ohio State Univ., Columbus, Ohio*, 1971) 233–257. MR0343355

RUPPERT, D. (1985). A Newton–Raphson version of the multivariate Robbins–Monro procedure. *Ann. Statist.* **13** 236–245. MR0773164 https://doi.org/10.1214/aos/1176346589

RUPPERT, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis* (B. K. Ghosh and P. K. Sen, eds.). *Statist. Textbooks Monogr.* **118** 503–529. Dekker, New York. MR1174318

SAKRISON, D. J. (1967). The use of stochastic approximation to solve the system identification problem. *IEEE Trans. Automat. Control* **12** 563–567.

SARIDIS, G. and STEIN, G. (1968). A new algorithm for linear system identification. *IEEE Trans. Automat. Control* **13** 592–594.

SIEGMUND, D. (2003). Herbert Robbins and sequential analysis. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 349–365. MR1983532 https://doi.org/10.1214/aos/1051027870

SPALL, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control* **37** 332–341. MR1148715 https://doi.org/10.1109/9.119632

SPALL, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Automat. Control* **45** 1839–1853. MR1795352 https://doi.org/10.1109/TAC.2000.880982

SPALL, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley, Hoboken, NJ. MR1968388 https://doi.org/10.1002/0471722138

TEMLYAKOV, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12** 213–227. MR1745113 https://doi.org/10.1023/A:1018917218956

TEWARI, A. and MURPHY, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health* (J. Re, S. A. Murphy and S. Kumar, eds.) 495–517. Springer, Berlin.

TOMKINS, S., LIAO, P., YEUNG, S., KLASNJA, P. and MURPHY, S. A. (2019). Intelligent pooling in Thompson sampling for rapid personalization in mobile health. ICML 2019 Workshop RL4RealLife, 2019.

WEI, C. Z. (1987). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15** 1115–1130. MR0902249 https://doi.org/10.1214/aos/1176350496

WIDROW, B. and HOFF, M. E. (1960). Adaptive switching circuits. In *Proc. IRE WESCON Convention Record, Part* 4 96–104.

WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186 https://doi.org/10.1214/16-EJS1108

YUAN, Q. and YIN, G. (2015). Analyzing convergence and rates of convergence of particle swarm optimization algorithms using stochastic approximation methods. *IEEE Trans. Automat. Control* **60** 1760–1773. MR3365066 https://doi.org/10.1109/TAC.2014.2381454

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 379–390. MR1983534

Statistical Science
2021, Vol. 36, No. 2, 303–327
https://doi.org/10.1214/20-STS785

Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning

Jianqing Fan, Kaizheng Wang, Yiqiao Zhong and Ziwei Zhu

Abstract. Factor models are a class of powerful statistical models that have been widely used to deal with dependent measurements that arise frequently from various applications from genomics and neuroscience to economics and finance. As data are collected at an ever-growing scale, statistical machine learning faces some new challenges: high dimensionality, strong dependence among observed variables, heavy-tailed variables and heterogeneity. High-dimensional robust factor analysis serves as a powerful toolkit to conquer these challenges.

This paper gives a selective overview on recent advance on high-dimensional factor models and their applications to statistics including Factor-Adjusted Robust Model selection (FarmSelect) and Factor-Adjusted Robust Multiple testing (FarmTest). We show that classical methods, especially principal component analysis (PCA), can be tailored to many new problems and provide powerful tools for statistical estimation and inference. We highlight PCA and its connections to matrix perturbation theory, robust statistics, random projection, false discovery rate, etc., and illustrate through several applications how insights from these fields yield solutions to modern challenges. We also present far-reaching connections between factor models and popular statistical learning problems, including network analysis and low-rank matrix recovery.

Key words and phrases: Factor model, PCA, covariance estimation, perturbation bounds, robustness, random sketch, FarmSelect, FarmTest.

REFERENCES

ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334 https://doi.org/10.1109/FOCS.2015.47

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. Preprint. Available at arXiv:1709.09565.

AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227. MR3064065 https://doi.org/10.3982/ECTA8968

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. MR0947576 https://doi.org/10.1214/aos/1176350834

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance ma-

Jianqing Fan, Fredreick L. Moore'18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, 08540, New Jersey, USA (e-mail: jqfan@princeton.edu). Kaizheng Wang, Assistant Professor, Department of Industrial Engineering and Operations Research, Columbia University, New York, 10027, New York, USA (e-mail: kw2934@columbia.edu). Yiqiao Zhong, Postdoctoral Research Fellow, Packard 239, Stanford University, Stanford, 94305, California, USA (e-mail: yiqiaoz@stanford.edu). Ziwei Zhu, Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, 48109, Michigan, USA (e-mail: ziweiz@umich.edu).

trices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

BARTLETT, M. S. (1938). Methods of estimating mental factors. *Nature* **141** 609–610.

BARTLETT, M. S. (1950). Tests of significance in factor analysis. *Br. J. Math. Stat. Psychol.* **3** 77–85.

BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000 https://doi.org/10.1145/1970392.1970395

CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Ann. Statist.* **47** 2405–2439. MR3988761 https://doi.org/10.1214/18-AOS1752

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

CATTELL, R. B. (1966). The scree test for the number of factors. *Multivar. Behav. Res.* **1** 245–276.

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050 https://doi.org/10.2307/1912275

COHEN, M. B., NELSON, J. and WOODRUFF, D. P. (2016). Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming. LIPIcs. Leibniz Int. Proc. Inform.* **55** Art. No. 11, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3577072

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001

DESAI, K. H. and STOREY, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *J. Amer. Statist. Assoc.* **107** 135–151. MR2949347 https://doi.org/10.1080/01621459.2011.645777

DOBRIBAN, E. (2017). Factor selection by permutation. Preprint. Available at arXiv:1710.00479.

DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302 https://doi.org/10.1198/016214506000001211

EFRON, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* **105** 1042–1055. MR2752597 https://doi.org/10.1198/jasa.2010.tm09129

ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: Spectral analysis beyond Davis–Kahan. In *Algorithmic Learning Theory* 2018. *Proc. Mach. Learn. Res.* (*PMLR*) **83** 38. Proceedings of Machine Learning Research PMLR. MR3857310

FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991 https://doi.org/10.1016/j.jeconom.2008.09.017

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. MR3689312 https://doi.org/10.1111/rssb.12204

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

FAN, J., KE, Y. and WANG, K. (2016a). Decorrelation of covariates for high dimensional sparse regression. Preprint. Available at arXiv:1612.08490.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

FAN, J., LIU, H. and WANG, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46** 1383–1414. MR3819104 https://doi.org/10.1214/17-AOS1588

FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** Paper No. 207, 42. MR3827095

FAN, J., WANG, W. and ZHU, Z. (2016b). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. Preprint. Available at arXiv:1603.08315.

FAN, J., KE, Y., SUN, Q. and ZHOU, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* **114** 1880–1893. MR4047307 https://doi.org/10.1080/01621459.2018.1527700

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

HIRZEL, A. H., HAUSSER, J., CHESSEL, D. and PERRIN, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* **83** 2027–2036.

HOCHREITER, S., CLEVERT, D.-A. and OBERMAYER, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics* **22** 943–949.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

HORN, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* **30** 179–185.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417.

HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *ITCS'*13*—Proceedings of the* 2013 *ACM Conference on Innovations in Theoretical Computer Science* 11–19. ACM, New York. MR3385380

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

JOLLIFFE, I. T. (1986). *Principal Component Analysis*. *Springer Series in Statistics*. Springer, New York. MR0841268 https://doi.org/10.1007/978-1-4757-1904-8

KENDALL, M. G. (1957). *A Course in Multivariate Analysis*. *Griffin's Statistical Monographs & Courses*, *No*. 2. Hafner Publishing Company, New York. MR0092297

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

KNEIP, A. and SARDA, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *Ann. Statist.* **39** 2410–2447. MR2906873 https://doi.org/10.1214/11-AOS905

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII*. *Progress in Probability* **71** 397–423. Springer, Cham. MR3565274 https://doi.org/10.1007/978-3-319-40519-3_18

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663 https://doi.org/10.1214/12-AOS970

LAWLEY, D. and MAXWELL, A. (1962). Factor analysis as a statistical method. *J. Roy. Statist. Soc. Ser. D* **12** 209–229.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

LI, Q., CHENG, G., FAN, J. and WANG, Y. (2018). Embracing the blessing of dimensionality in factor models. *J. Amer. Statist. Assoc.* **113** 380–389. MR3803472 https://doi.org/10.1080/01621459.2016.1256815

MCCRAE, R. R. and JOHN, O. P. (1992). An introduction to the five-factor model and its applications. *J Pers.* **60** 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

MOR-YOSEF, L. and AVRON, H. (2019). Sketching for principal component regression. *SIAM J. Matrix Anal. Appl.* **40** 454–485. MR3945241 https://doi.org/10.1137/18M1188860

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*.

O'ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: A survey. *J. Combin. Theory Ser. A* **144** 361–442. MR3534074 https://doi.org/10.1016/j.jcta.2016.06.008

O'ROURKE, S., VU, V. and WANG, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra Appl.* **540** 26–59. MR3739989 https://doi.org/10.1016/j.laa.2017.11.014

ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **92** 1004–1016.

ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766 https://doi.org/10.1016/j.jeconom.2012.01.034

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PAUL, D., BAIR, E., HASTIE, T. and TIBSHIRANI, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.* **36** 1595–1618. MR2435449 https://doi.org/10.1214/009053607000000578

PAULSEN, V. (2002). *Completely Bounded Maps and Operator Algebras*. *Cambridge Studies in Advanced Mathematics* **78**. Cambridge Univ. Press, Cambridge. MR1976867

PEARSON, K. (1901). Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal* **6** 566.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*.

SHKOLNISKY, Y. and SINGER, A. (2012). Viewing direction estimation in cryo-EM using synchronization. *SIAM J. Imaging Sci.* **5** 1088–1110. MR3022188 https://doi.org/10.1137/120863642

SPEARMAN, C. (1927). The abilities of man.

SRIVASTAVA, N. and VERSHYNIN, R. (2013). Covariance estimation for distributions with $2+\varepsilon$ moments. *Ann. Probab.* **41** 3081–3111. MR3127875 https://doi.org/10.1214/12-AOP760

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. *Computer Science and Scientific Computing*. Academic Press, Boston, MA. MR1061154

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271 https://doi.org/10.1198/016214502388618960

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TRON, R. and VIDAL, R. (2009). Distributed image-based 3-D localization of camera sensor networks. In *Decision and Control*, 2009 *Held Jointly with the* 2009 28*th Chinese Control Conference. CDC/CCC* 2009. *Proceedings of the* 48*th IEEE Conference on* IEEE, New York.

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found*. *Comput*. *Math*. **12** 389–434. MR2946459 https://doi.org/10.1007/s10208-011-9099-z

VERSHYNIN, R. (2012a). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

VERSHYNIN, R. (2012b). How close is the sample covariance matrix to the actual covariance matrix? *J. Theoret*. *Probab*. **25** 655–686. MR2956207 https://doi.org/10.1007/s10959-010-0338-z

WANG, H. (2012). Factor profiled sure independence screening. *Biometrika* **99** 15–28. MR2899660 https://doi.org/10.1093/biomet/asr074

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann*. *Statist*. **45** 1342–1374. MR3662457 https://doi.org/10.1214/16-AOS1487

WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann*. *Statist*. **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968 https://doi.org/10.1007/bf01932678

WOODRUFF, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found*. *Trends Theor. Comput*. *Sci*. **10** iv+157. MR3285427 https://doi.org/10.1561/0400000060

YANG, J., MENG, X. and MAHONEY, M. W. (2016). Implementing randomized matrix algorithms in parallel and distributed environments. *Proc*. *IEEE* **104** 58–92.

YI, X., CARAMANIS, C. and SANGHAVI, S. (2016). Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. Preprint. Available at arXiv:1608.05749.

YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006 https://doi.org/10.1093/biomet/asv008

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J*. *Mach*. *Learn*. *Res*. **7** 2541–2563. MR2274449

ZHONG, Y. (2017). Eigenvector under random perturbation: A nonasymptotic Rayleigh–Schrö dinger theory. Preprint. Available at arXiv:1702.00139.

ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM J*. *Optim*. **28** 989–1016. MR3782406 https://doi.org/10.1137/17M1122025

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

Statistical Science
2021, Vol. 36, No. 2, 328–337
https://doi.org/10.1214/20-STS801

A Conversation with Dennis Cook

Efstathia Bura, Bing Li, Lexin Li, Christopher Nachtsheim, Daniel Pena, Claude Setodji and Robert E. Weiss

Abstract. Dennis Cook is a Full Professor, School of Statistics, at the University of Minnesota. He received his BS degree in Mathematics from Northern Montana College, and MS and PhD degrees in Statistics from Kansas State University. He has served as Chair of the Department of Applied Statistics, Director of the Statistical Center and Director of the School of Statistics, all at the University of Minnesota.

His research areas include dimension reduction, linear and nonlinear regression, experimental design, statistical diagnostics, statistical graphics and population genetics. He has authored over 200 research articles and is author or co-author of two textbooks— An Introduction to Regression Graphics and Applied Regression Including Computing and Graphics—and three research monographs, Influence and Residuals in Regression, Regression Graphics: Ideas for Studying Regressions through Graphics and An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics.

He has served as Associate Editor of the *Journal of the American Statistical Association, The Journal of Quality Technology, Biometrika, Journal of the Royal Statistical Society* and *Statistica Sinica*. He is a four-time recipient of the Jack Youden Prize for Best Expository Paper in *Technometrics* as well as the Frank Wilcoxon Award for Best Technical Paper. He received the 2005 COPSS Fisher Lecture and Award, and he is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics. The following conversation took place on March 22, 2019, following the banquet at the conference, "Cook's Distance and Beyond: A Conference Celebrating the Contributions of R. Dennis Cook." The interviewers were, Efstathia Bura (Effie), Bing Li, Lexin Li, Christopher Nachtsheim (Chris), Daniel Pena, Claude Messan Setodji and Robert Weiss (Rob).

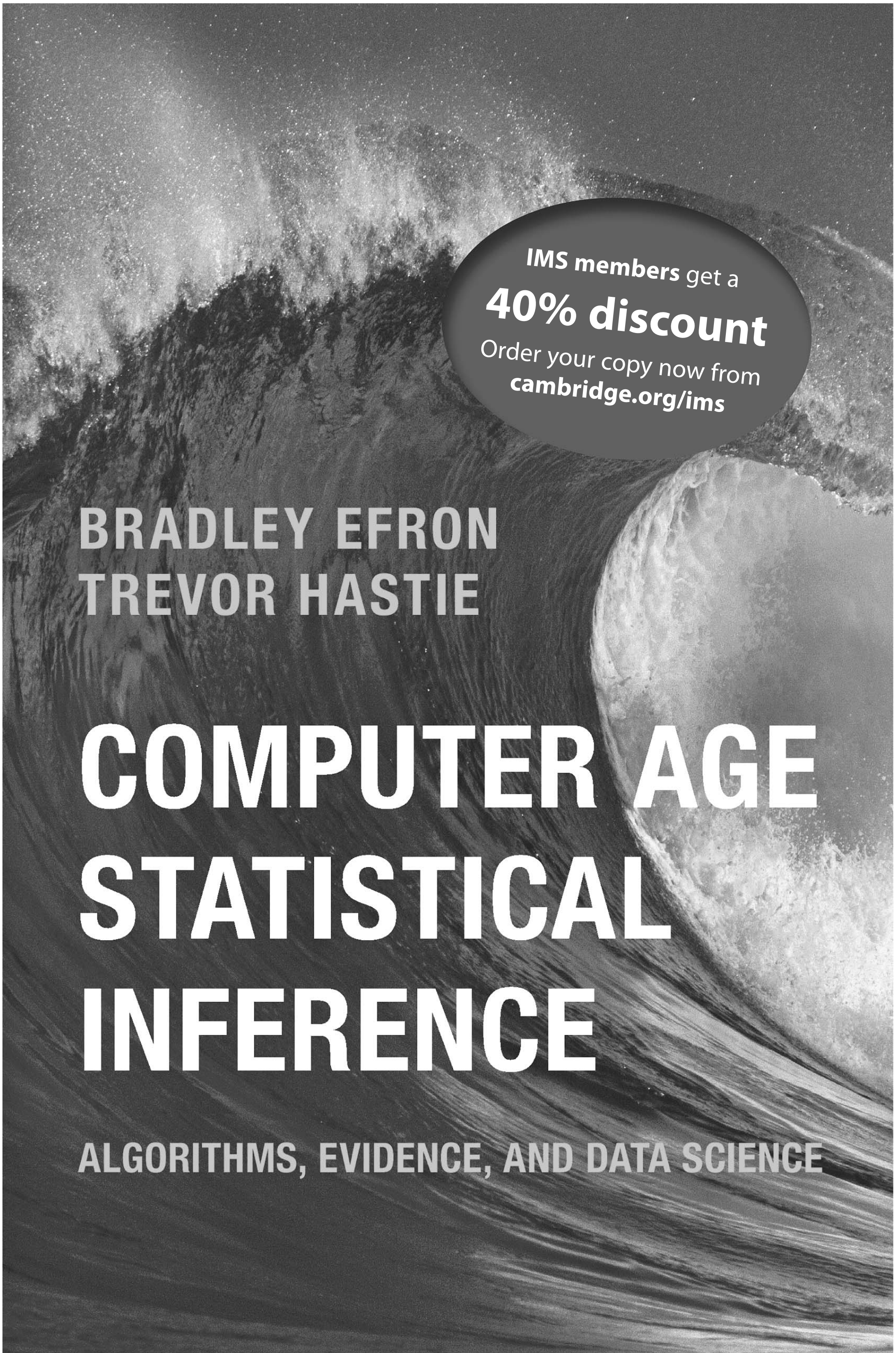
Key words and phrases: Bayesian methods, central subspace, Cook's distance, envelope methods, influence, neural networks, optimal experimental design, regression diagnostics, sufficient dimension reduction.

Efstathia Bura is University Professor and Chair of the Applied Statistics Research Group, Institute of Statistics and Mathematical Methods in Economics, Faculty of Mathematics and Geoinformation, TU Wien, Vienna, Austria (e-mail: Efstathia.bura@tuwien.ac.at). Bing Li is Verne M. Willaman Professor of Statistics, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, 16802, USA (e-mail: bxl9@psu.edu). Lexin Li is Professor, Department of Biostatistics and Epidemiology, and Helen Wills Neuroscience Institute, University of California, Berkeley, California, 94720, USA (e-mail: lexinli@berkeley.edu). Christopher Nachtsheim is the Frank A. Donaldson Chair of Operations Management, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota, 55455, USA (e-mail: nacht001@umn.edu). Daniel Pena is Professor of Statistics, Institute of Big Data and Department of Statistics, Universidad Carlos III de Madrid, Madrid 28903, Spain (e-mail: daniel.pena@uc3m.es). Claude Setodji is the co-director of the RAND Center for Causal Inference, RAND Corporation, Pittsburgh, Pennsylvania, 15231, USA (e-mail: setodji@rand.org). Robert E. Weiss is Professor, Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, 90095, USA (e-mail: robweiss@ucla.edu).

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.


IMS members get a
40% discount
Order your copy now from
cambridge.org/ims
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE
STATISTICAL
INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE