Technical Surprises

My surprise at learning that the "mean of a multivariate normal distribution is not best estimated by the sample mean" was genuine but, apparently, not unique. The result, which follows from the development of the James–Stein estimator (JSE), has apparently been shocking people since it was published in 1961 [1], building on prior work by Stein in 1956 [2].

Reference [3] gives an easily accessible discussion of this apparent paradox in the context of baseball statistics. For a more technical result, consider the estimation of a normally distributed random variable with p elements, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 I)$, with σ assumed known. It is well known that the maximum-likelihood estimate (MLE) is given by $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{X}$, with a standard figure of merit of the estimate being the mean-squared error (MSE)

$$J(\boldsymbol{\theta}) = E(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2),$$

which, in this case, yields $J^{\mathrm{ML}}(\boldsymbol{\theta}) = p\sigma^2$.

The ML estimator of $\boldsymbol{\theta}$ was thought to be the best available, but James and Stein demonstrated that their filter dominates the MLE in that

$$J^{\mathrm{JS}}(\boldsymbol{\theta}) < J^{\mathrm{ML}}(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}, \text{ provided } p > 2.$$

This observation apparently led to "long periods of resistance … punctuated by frequent and angry debate" [3], but the results are more broadly accepted now.

Digital Object Identifier 10.1109/MCS.2016.2584381
Date of publication: 16 September 2016

The result, which follows from the development of the James–Stein estimator, has apparently been shocking people since it was published in 1961.

The form of the JSE for this particular application is

$$\hat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{\sigma^2(p-2)}{\mathbf{X}^T\mathbf{X}}\right)\mathbf{X},$$

which is a special type of "shrinkage estimator" in that the second term on the right shrinks the MLE $\mathbf{X}$ toward some centralized mean [4] (and note that this shrinkage can be negative, which makes the JSE itself inadmissible). The form of the JSE leads to several interesting observations, including the interpretation that the JSE is essentially an *empirical Bayes estimator* [5, p. 273] [6], which has significant implications in terms of the Bayesian robustness of the JSE when the assumptions about the hyperparameters of the prior are incorrect. It is important to note that $\hat{\boldsymbol{\theta}}_{JS}$ is a biased, nonlinear estimator, but it is the property ($J^{JS}(\boldsymbol{\theta}) < J^{\mathrm{ML}}(\boldsymbol{\theta})$) that might make it particularly useful for applications that place high value on reduced uncertainty in the estimate. For further generality, [7, Theorem 7] and [4, p. 2434] discuss the extension of these results to the case where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}, Q)$, with $Q = Q^T > 0$.

In 1997–1998, [4] and [8] extended this prior work by developing a recursive form of the JSE for parameter estimation of autoregressive with exogenous input models and for state-space systems. The main results are presented as the James-Stein state filter (JSSF), which is shown to preserve the performance benefits discussed previously (in fact it is shown that $J_k^{\mathrm{JSSF}} < J_k^{\mathrm{ML}}$, with the MLE in this case only using the observations) and have similar computational complexity to the Kalman filter. A remarkable result of [4] is that these performance benefits hold regardless of the inaccuracies in the propagation dynamics, including the possibility of perturbations to the A and B matrices in the dynamical system, $\mathbf{x}_{k+1} = A\mathbf{x}_k + B\mathbf{e}_k$, and nonnormality of the noise $\mathbf{e}$. Furthermore, it is shown that "the JSSF is a globally robust state filter" in that "for sufficiently large modeling errors, the JSSF is expected to outperform any locally robust Kalman filter simply because the JSSF has a global upper bound on its mean-square error" [4]. Examples of locally robust Kalman filters that the authors discuss include the heavily cited papers [9] and [10], although there

This editorial was submitted before we learned the sad news of the passing of Rudolf Kalman on July 2, 2016. He was a giant in the field and will certainly be missed. Actions are already underway to publish an obituary in *IEEE Control Systems Magazine* and also to possibly dedicate a future issue to recognize his legacy.

Jonathan How with Siva Banda at the AIAA Fellows Dinner, 2016.

are possibly more recent approaches that could yield better results.

These results are not without their own limitations, which [4, p. 2443] does a good job of discussing. The main issue is that, while the JSE improves the overall MSE, it does not necessarily improve the MSE of each element of $\mathbf{X}$. Furthermore, the JSSF assumes that the observation matrix has at least as many rows as columns, which might require reduced-order, state-space models in some applications.

Being immersed in a community that celebrates the significant successes of the Kalman filter, but at the same time laments its divergence issues [11]–[13], these JSE results were surprising because I was not aware of this ongoing discussion in the signal-processing community. However, with only approximately 37 citations (and only one citation of the related paper presented at the 1997 Conference on Decision and Control [8]), it is not clear how well known these filter results are in either the signal-processing or control-systems communities.

With the limitations that have been provided, the JSSF may not be the answer to the often-posed question "what is next after the Kalman filter?," but the results seem interesting enough to merit a deeper look for applications that require a robust filter and/or could provide a launching point for future research into algorithms that relax some of the assumptions made in the JSSF while retaining most of the benefits. And if not that, then hopefully at least this discussion provides you with something to stump your colleagues on at the next coffee hour!

Please let me know of any technical surprises that you have discovered.

REFERENCES

[1] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, 1961, vol. 1, pp. 361–379.

[2] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, 1956, vol. 1, pp. 197–206.

[3] B. Efron and C. Morris, "Stein's paradox in statistics," *Sci. Amer.*, vol. 236, pp. 119–127, May 1977.

[4] J. H. Manton, V. Krishnamurthy, and H. Vincent Poor, "James-Stsssein state filtering algorithms," *IEEE Trans. Signal Processing*, vol. 46, no. 9, pp. 2431–2447, 1998.

[5] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1991.

[6] B. Efron and C. Morris, "Stein's estimation rule and its competitors: An empirical Bayes approach," *J. Am. Stat. Assoc.*, vol. 68, no. 341, pp. 117–130, 1973.

[7] M. E. Bock, "Minimax estimators of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 3, no. 1, pp. 209–218, 1975.

[8] J. H. Manton, V. Krishmamurthy, and H. Vincent Poor, "James-Stein state space filter," in *Proc. 36th IEEE Conf. Decision and Control*, 1997, vol. 4, pp. 3454–3459.

[9] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[10] L. Xie, Y. C. Soh, and C. E. de Souza, "Robust Kalman filtering for uncertain discrete-time systems," *IEEE Trans. Automat. Control*, vol. 39, no. 6, pp. 1310–1314, 1994.

[11] R. J. Fitzgerald, "Divergence of the Kalman filter," *IEEE Trans. Automat. Control*, vol. 16, no. 6, pp. 736–747, 1971.

[12] F. H. Schlee, C. J. Standish, and N. F. Toda, "Divergence in the Kalman filter," *AIAA J.*, vol. 5, no. 6, pp. 1114–1120, 1967.

[13] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

Jonathan P. How

Historical Controversy

The method of least squares is the automobile of modern statistical analysis: despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all. But there has been some dispute, historically, as to who was the Henry Ford of statistics. Adrien Marie Legendre published the method in 1805, an American, Robert Adrain, published the method in late 1808 or early 1809, and Carl Friedrich Gauss published the method in 1809. Legendre appears to have discovered the method in early 1805, and Robert Adrain may have "discovered" it in Legendre's 1805 book, but in 1809 Gauss had the temerity to claim that he had been using the method since 1795, and one of the most famous priority disputes in the history of science was off and running.

—Stephen M. Stigler, "Gauss and the Invention of Least Squares," *The Annals of Statistics*, vol. 9, no. 3, pp. 465–474, 1981.