Approximate repeated-measures shrinkage

Adam R. Brentnall[a 1], Martin J. Crowder[b], David J. Hand[b,c]

a Wolfson Institute for Preventive Medicine, Queen Mary University of London.

b Institute for Mathematical Sciences, Imperial College London.

c Department of Mathematics, Imperial College London.

Abstract

A general method is formalised for the problem of making predictions for a fixed group of individual units, following a sequence of repeated measures on each. A review of some related work is undertaken and, using some of its terminology, the approach might be described as approximate non-parametric empirical Bayes prediction. It is contended that the method may often produce predictions that are, in practice, comparable or not much worse than more sophisticated methods, but sometimes for a smaller computational cost. Two examples are used to demonstrate the approach, exploring the prediction of baseball averages and spatial-temporal rainfall. The method performs favourably in both examples in comparison with James-Stein, empirical Bayes and other predictions; it also provides a relatively simple and computationally feasible way of determining whether it is worth modelling between-individual variability.

Keywords: Empirical Bayes, Prediction, Random Effects

[1] Corresponding Author: Cancer Research UK Centre for Epidemiology, Mathematics and Statistics, Wolfson Institute for Preventive Medicine, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK. Email: a.brentnall@qmul.ac.uk. Tel: +44 (0)20 7882 3531, Fax: +44 (0)20 7882 3890.

1. Introduction

The problem considered in this paper is prediction of individual behaviour, when repeated measurements are observed on a large number of individual units. Specifically, if $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ are the data observed on individual $i = 1, \ldots, n$, then the objective is to predict $y_{i,n_i+1}, \ldots$ for each i. Databases are used to record repeated measurements on units such as consumers, sport stars, bank accounts or households, in a range of areas, from personal finance to cricket to supermarket transactions. This motivates methods to predict individual behaviour for large fixed n, but where n_i may vary between individuals.

Predictions about individual i that only use $\boldsymbol{y}_i$ can be improved upon by using the whole data $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. Such predictions follow from supposing that, given $\boldsymbol{U}_n = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, with $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{im})$, the $\boldsymbol{y}_i$'s are independent with probability densities or mass functions $p(\boldsymbol{y}_i|\boldsymbol{u}_i)$, where the $\boldsymbol{u}_i$ are independent random effects drawn from some common distribution. Note that although the $\boldsymbol{y}_i$'s are independent given $\boldsymbol{U}_n$, the components of $\boldsymbol{y}_i$ need not necessarily be independent given $\boldsymbol{u}_i$. This structure usually results in the estimates, or predictions of $\boldsymbol{u}_i$, being shrunk towards their sample mean. In this work we compare commonly-used shrinkage estimates against the following approach, called ARMS (approximate repeated-measures shrinkage).

The $\boldsymbol{u}_i$ $(i = 1, \ldots, n)$ are possibly drawn from some wider population about which we make no assumptions. We focus on the sampled $\boldsymbol{u}_i$ and place equal weight on them. In effect, our finite population of interest is $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ with probabilities assigned uniformly as $1/n$; we do not necessarily assume that the $\boldsymbol{u}_i$ are all different. Unlike the usual case, it is the support points $\boldsymbol{u}_i$ that are to be estimated rather than the distribution of probabilities. Let $\hat{\boldsymbol{u}}_i$ be a consistent estimate for $\boldsymbol{u}_i$, a maximum likelihood estimate for instance, and let $\pi_e(\boldsymbol{u})$ be the corresponding empirical distribution, assigning probability $1/n$ to each $\hat{u}_i$. In the following $p_e(.)$ and $P_e(.)$ will be used for probabilities based on this empirical distribution, that is, based on the $\hat{\boldsymbol{u}}_i$ rather than on the $\boldsymbol{u}_i$. Then, following the law of total probability, we construct predictions on random variables or events C of interest as

$$p_e(C|\boldsymbol{y}_i) \quad = \quad \sum_{j=1}^{n} p(C, \boldsymbol{u}_i = \hat{\boldsymbol{u}}_j | \boldsymbol{y}_i). \tag{1}$$

Thus the predictions are a summation

$$p_e(C|\boldsymbol{y}_i) \quad = \quad \sum_{j=1}^{n} z_{ij} p(C|\boldsymbol{y}_i, \hat{\boldsymbol{u}}_j). \tag{2}$$

with weights z_{ij} obtained by Bayes' rule

$$\begin{aligned} z_{ij} &= P_e(\boldsymbol{u}_i = \hat{\boldsymbol{u}}_j | \boldsymbol{y}_i) \\ &= \frac{p_e(\boldsymbol{y}_i|\hat{\boldsymbol{u}}_j)}{\sum_{k=1}^{n} p_e(\boldsymbol{y}_i|\hat{\boldsymbol{u}}_k)}. \end{aligned} \tag{3}$$

This approach has two desirable features. Firstly, for fixed n, as the n_i become large $(\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n) \rightarrow (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ in probability. More precisely, the ARMS estimate $p_e(.)$ converges to a distribution with mass $1/n$ on atoms $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ as $\min(n_i) \rightarrow \infty$, because each estimate $\hat{\boldsymbol{u}}_i$ is consistent. Secondly, computational requirements may be modest: estimation is only required for the $\hat{\boldsymbol{u}}_i$ (and iterative optimisation routines need not always be required for this), and the increase in computational complexity in going from univariate to multivariate random effects may be small. Once the individual estimates $\hat{\boldsymbol{u}}_i$ have been obtained, the computational complexity is $0(n^2)$ to calculate z_{ij}.

In some other shrinkage methods the random-effects distribution $\pi(\boldsymbol{u})$ is first estimated, and then used as a prior in Bayes formula to obtain posterior $\boldsymbol{u}_i$ quantities. This has sometimes been called *empirical Bayes* [11]. Estimation is a mixture problem, with likelihood

$$p(\boldsymbol{Y}_n;\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{n} \int p(\boldsymbol{y}_i|\boldsymbol{u}_i)p(\boldsymbol{u}_i;\boldsymbol{\theta})d\boldsymbol{u}_i \qquad (4)$$

where $\boldsymbol{\theta}$ are the parameters to be estimated. This setup assumes that there are no fixed effects in the model $p(\boldsymbol{y}_i|\boldsymbol{u}_i)$. If a model is specified for $p(\boldsymbol{u}_i;\boldsymbol{\theta})$ then some have termed this *parametric empirical Bayes* [25]. An alternative is to leave the form of $p(\boldsymbol{u}_i;\boldsymbol{\theta})$ unspecified, but to allow it a number of atoms at unknown points and mass in the $\boldsymbol{u}$ space, and this has been called *nonparametric maximum likelihood estimation* [18]. A further possibility is to take $p(\boldsymbol{u}_i;\boldsymbol{\theta})$ to be generated by a Dirichlet process [24], which fits into a class of techniques called *nonparametric Bayes* [26]. Finally, one might take a *parametric Bayesian approach* and simulate directly from the posterior distributions.

It can be seen that a range of different methods is available. However, most are likely to be more difficult to apply for large data sets and multivariate random effects than ARMS. A reason for this is that it may not be possible to write the integral in equation (4) as an analytical expression, and so one evaluation of the overall likelihood function will involve n separate numerical evaluations of a multi-dimensional integral. One aim of this article is to review the motivation and properties of different shrinkage methods, in order to help understand what may be lost by using the computationally more attractive ARMS.

The outline for the rest of the paper is as follows. In Section 2 a brief history of shrinkage estimators and the terminology introduced is set out. Then in Section 3 the different approaches are compared in more detail. The later sections contrast the performance of the proposed ARMS approach on examples reported previously in the literature, and demonstrate its application to a large data set.

2. Shrinkage estimators

2.1. A selective history

Early work on shrinkage estimation was motivated by data $y_1, y_2, \dots, y_n$ as realisations of independent random variables with parameters $u_1, u_2, \dots, u_n$. The objective is to estimate $u_1, \dots, u_n$ given a performance measure. Particular attention has been paid to two special cases.

The first case is when the data are normal random variables with unknown mean parameters. Stein [32] showed that estimates $\hat{\boldsymbol{U}}_n = \boldsymbol{Y}_n$ will not minimise the sum of squared errors $\mathrm{E}(|\boldsymbol{U}_n - \hat{\boldsymbol{U}}_n|^2)$, and James and Stein [17] derived a better estimate. This result has been called Stein's paradox, since some saw in it the possibility of combining apparently unrelated problems. Half a century later, we can see that the result has not been used in this way. To quote an example used in 1973 [12], the speed of light and the weight of hogs in Montana are not used to improve estimates of tea consumption in Taiwan. A more useful interpretation of the result is given by [33], who frames it using Galton's 'regression to mediocrity' idea.

The second case is a compound decision problem where $u_i = \pm 1$, and a decision is required on the sign of each so that the expected number of errors is minimised. Robbins [28] showed that a shrinkage estimator beats the simple approach of using the sign of the observations. He also suggested that for more general problems a decision rule could be formed from estimates of $\pi(\boldsymbol{u})$. Robbins called this approach empirical Bayes [29], missing the opportunity for a more catchy name (the idea had previously been described as "an attempt to lift ourselves by our own bootstraps" [28]). A distinction between empirical Bayes and compound decisions used to be made [10], but the differences are largely ignored today [11]. It can be shown that the James-Stein estimator is also an empirical Bayes estimator [25], so in some sense, Robbins anticipated

Stein's result.

Shrinkage estimates are now formed by many different statistical models. For example, Bayesian linear models were partly motivated by the above results [22], and ideas about the variance-bias estimation trade-off are applied by regularisation methods such as penalised likelihood. Two recent papers in this journal to use shrinkage estimates are [21] and [15].

2.2. On the proposed approach

The problems in Sections 1 and 2.1 are quite similar. The most obvious difference is that in the first one the data and parameters are allowed to be multi-dimensional. But, there is also another difference that led Robbins to discard the ARMS prediction method [28].

Robbins considered using an empirical distribution of the observed data as an estimate of $\pi(\boldsymbol{u})$, but found that the estimate might be biased, with bias not tending to zero as $n \to \infty$ ([28], p. 142). This is important for the compound decision problem because the observations from the random variables continue to arrive [30].

However, the potential for such a bias may be less relevant for the problem described in Section 1 because the number of individual units is taken to be fixed. The finite population of interest means that while there is theoretical interest in $n_i \to \infty$ with n fixed, there is none here in $n \to \infty$ with n_i fixed. Thus, the use of an empirical distribution based on the observed data to derive the ARMS prediction equation (2) may still have some merit in many real situations.

3. Estimating the random-effects distribution

The ARMS approach might be seen under an empirical Bayes umbrella, as another way to estimate $\pi(\boldsymbol{u})$. In this section we explore how it compares to other estimation procedures.

3.1. Nonparametric models

The nonparametric maximum likelihood (NPML) estimate is a discrete distribution, with number of mass points $k \leq n$ [18]. One effect of this is that the integral in equation (4) is a summation, simplifying computation. The NPML estimate is not necessarily the same as the ARMS estimate, indeed the approach in Section 1 is called 'approximate' with regard to NPML, *i.e.* ARMS predictions approximate those from NPML estimation. However, as $\min(n_i) \to \infty$ for n fixed they should converge to the same distribution (but of course not when $n \to \infty$ with n_i fixed). This is true for the following reason. Firstly, the weights $\pi(u_i) = 1/n$ for $i = 1, \ldots, n$ as $\min(n_i) \to \infty$ by definition. Secondly, as the $n_i \to \infty$, $p(\boldsymbol{y}_i|\boldsymbol{u}_j)/p(\boldsymbol{y}_i|\boldsymbol{u}_i) \to 0$ for $\boldsymbol{u}_j \neq \boldsymbol{u}_i$ and, therefore, the $\hat{\boldsymbol{u}}_i$ (which $\to \boldsymbol{u}_i$) will eventually maximise all n sums and hence the likelihood.

Several algorithms have been proposed for NPML estimation, including EM and intra-simplex directional methods [20] and alternatives are still being developed, *e.g.* [34]. A practical issue is that computational complexity may sometimes increase exponentially with dimension of $\boldsymbol{u}_i$, because the number of parameters for the location of each mass point is equal to the dimension of $\boldsymbol{u}_i$.

A continuous distribution might be preferred to a discrete distribution if interest is in making predictions for new individuals. That is, when the underlying $\pi(\boldsymbol{u})$ really is continuous. [19] consider a computationally efficient way to smooth non-parametric maximum likelihood estimates, but also note that NPML may still be competitive against correctly specified parametric distributions.

Another approach to smoothing the NPML estimate is to represent uncertainty on the distribution function $\pi(\boldsymbol{u})$ as a probability distribution on the space of distribution functions, that is as a random probability measure. A Dirichlet process is commonly used [24]. This approach can sometimes be computationally intensive even for scalar $\boldsymbol{u}_i$, but simulation algorithms, such as Gibbs sampling, can be used to fit [23].

3.2. Parametric models

It might be possible to choose a form $p(\boldsymbol{u};\boldsymbol{\theta})$ that provides an explicit form for the integral in equation (4). In this case, the parameters $\boldsymbol{\theta}$ might be estimated by maximising (4) with an unconstrained optimisation method, such as a quasi-Newton algorithm. Otherwise, the integral will need to be evaluated numerically. When Monte Carlo simulation is used and the likelihood is explored using an optimisation algorithm the approach is sometimes called simulated maximum likelihood; if an EM algorithm is used it has been called Monte Carlo EM [8]. Unfortunately, as n and the dimension of $\boldsymbol{u}_i$ increase, the approach rapidly becomes unfeasible.

3.3. Direct prediction

The original proposal by [29] did not estimate $\pi(\boldsymbol{u})$ and then to plug it in to Bayes' formula, but used an estimate for $p(\boldsymbol{y}_i) = \int p(\boldsymbol{y}_i|\boldsymbol{u})\pi(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ directly in the prediction equation for $\mathrm{E}(\boldsymbol{u}|\boldsymbol{y}_i)$. The best-known example of this approach from [29] is when $p(\boldsymbol{y}_i|\boldsymbol{u})$ is of Poisson form. The direct approach has sometimes been called non-parametric empirical bayes (NPEB).

The direct approach has been followed by others, including [7]. Brown [5] showed how the Bayes estimate can sometimes be obtained by shrinking the estimate based on the individual's data by a term $\{\partial p(\boldsymbol{y}_i)/\partial \boldsymbol{y}_i\}/p(\boldsymbol{y}_i)$. This approach was exploited in [6] by using a kernel estimator for $p(\boldsymbol{y}_i)$.

3.4. Comparison with ARMS

Many ways may be used to estimate $\pi(\boldsymbol{u})$, and it is clear that at least some of the above methods can perform better than $\pi_e(\boldsymbol{u})$. For example, if a binomial model is taken for the individual and half the individuals have only one observation ($n_i = 1$) and the other half have 100 then $\pi_e(\boldsymbol{u})$ might be a very poor approximation to $\pi(\boldsymbol{u})$: half of the mass (resulting from individuals with $n_i = 1$) will be shared between the values of 0 and 1. However, $\pi(\boldsymbol{u})$ is only a step on the way to prediction, and the numbers of repeated observations are implicitly considered through (3). For instance, those with 100 observations are unlikely to be shrunk much for a wide variety of different $\pi(\boldsymbol{u})$ estimates: Morris notes that the main benefit may be from the richer model form, rather than the choice of shrinkage technique [25]. Indeed, the direct approach is partly motivated by the observation that it may be more practical and effective not to estimate $\pi(\boldsymbol{u})$ [6]. We suggest that ARMS may produce shrinkage estimates that are not very different to those from more sophisticated and computationally-expensive methods. In the next section we investigate this claim through some empirical examples.

4. Baseball

There has been some interest shown in using shrinkage estimates to predict individual-player baseball averages, including by [25]. The baseball prediction problem fits into the general framework of Section 1. There are n individuals, and each has been observed n_i times, where the outcome $y_{ij} = 1$ if the ball is hit, 0 otherwise. The objective is to predict the proportion of balls hit in the future, for each individual. Note that the number of individuals to be predicted is fixed, and the n_i increase. In fact, the model and objective described in Section 1 provide a better description of the problem than the traditional compound-decision setup described in Section 2.

We next introduce the performance measures and shrinkage methods used, and then present the results.

4.1. Performance measures

Estimates of u_i are evaluated in [13] using the sum of squared differences $\sum_{i=1}^{n}(\hat{u}_i - o_i)^2$ with the observed proportion o_i of hits for the rest of the season. A performance measure is derived in [6] to more accurately measure the quantity of interest $\sum_{i=1}^{n}(\hat{u}_i - u_i)^2$. More precisely, setting the proportion of hits in the second half of the season r_i as the ratio of the number of hits $y_i^{(2)}$ and total at-bats $n_i^{(2)}$, then the measure is

$$\widehat{TSE}_R(\hat{\boldsymbol{u}}) \quad = \quad \sum_{i=1}^{n}(r_i - \hat{u}_i)^2 - \sum_{i=1}^{n} r_i(1-r_i)/n_i^{(2)} \tag{5}$$

where $\hat{\boldsymbol{u}}$ are the predictions, and $\widehat{TSE}_R$ stands for the total squared error for the proportion of hits R. We report $\widehat{TSE}^*_R(\hat{\boldsymbol{u}})$, which normalises $\widehat{TSE}_R(\hat{\boldsymbol{u}})$ to be be 1.0 for predictions using individual maximum likelihood estimates, *i.e.* the proportion of hits in the first three months.

An alternative performance measure is a Brier score [4], which is often used to rank forecasting systems [16]. When an individual i has outcome y_{ij} (0 or 1) and corresponding prediction $\hat{u}_i$ the overall average Brier score is $(\sum_{i=1}^{n} n_i)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\hat{u}_i - y_{ij})^2$. This seems (to us) to be a more sensible measure of overall model performance than the sum of squared differences $\sum_{i=1}^{n}(\hat{u}_i - o_i)^2$, or related quantities, because the Brier score weights each prediction equally, whereas the sum of squares implicitly downweights the individual predictions from the people with more observations.

To obtain standard errors on the performance measures, a non-parametric bootstrap was applied using 1000 bootstraps, resampling each individual's fitting and validation sets from their complete season.

4.2. Shrinkage methods

Individual MLEs, an overall mean and ARMS predictions were used. Note that the ARMS approach may be applied without modification when n_i vary for $i = 1, \ldots, n$, because the number of observations is automatically considered in the z_{ij} term in equation (2) via the likelihood $p(\boldsymbol{y}_i|\boldsymbol{u})$. For NPML predictions we used an EM algorithm on untransformed data and checked convergence using a stopping rule, and double-checked convergence by calculating the directional derivative [34].

The Stein prediction approaches from [13] and [6] were taken. These stabilise the variance of observed proportions $q_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ through a transformation. An arc-sine transformation $w_i = \arcsin(q_i)$ was used in [13], with variance $v_i = (4n_i)^{-1}$. This presentation of the transformation from [6] is slightly different from that presented in [13], because it also caters for the case where the n_i vary. [6] recommended an alternative transformation $x_i = \arcsin\sqrt{(\sum_{j=1}^{n_i} y_{ij} + 0.25)/(n_i + 0.5)}$ because it has better asymptotic control over bias. Stein predictions of the transformed variables are as [6], their equation (4.17). Predictions using the w transformation are denoted Stein 1, those from the x transformation are Stein 2. We also applied the NPEB method described in [6], using an x transformation.

4.3. Batting averages in 1970

The data and predictions are presented in Table 1. The last column in the table records the proportion of balls hit in the rest of the season. All the methods shrink predictions relative to individual MLEs, the NPML predictions being shrunk the most, but the differences are small. Table 2 presents the results. They are in line with [13]: the differences are small between the shrinkage methods. Further inspection of the performance measures for each individual $i = 1, \ldots, 18$ show that no single shrinkage method is consistently lower than

Table 1: Baseball data and predictions. Each player is observed 45 times before predictions are made.

Player	# Suc	MLE	ARMS	Stein 1	Stein 2	NPEB	NPML	Outcome
1	18	0.400	0.344	0.290	0.290	0.309	0.285	0.346
2	17	0.378	0.333	0.286	0.286	0.324	0.281	0.298
3	16	0.356	0.320	0.281	0.282	0.323	0.276	0.276
4	15	0.333	0.306	0.277	0.278	0.314	0.273	0.222
5	14	0.311	0.291	0.273	0.274	0.300	0.269	0.273
6	14	0.311	0.291	0.273	0.274	0.300	0.269	0.270
7	13	0.289	0.275	0.268	0.270	0.269	0.266	0.263
8	12	0.267	0.261	0.264	0.266	0.230	0.264	0.210
9	11	0.244	0.248	0.259	0.262	0.220	0.262	0.269
10	11	0.244	0.248	0.259	0.262	0.220	0.262	0.230
11	10	0.222	0.237	0.254	0.258	0.238	0.260	0.264
12	10	0.222	0.237	0.254	0.258	0.238	0.260	0.256
13	10	0.222	0.237	0.254	0.258	0.238	0.260	0.303
14	10	0.222	0.237	0.254	0.258	0.238	0.260	0.264
15	10	0.222	0.237	0.254	0.258	0.238	0.260	0.226
16	9	0.200	0.227	0.249	0.253	0.261	0.259	0.285
17	8	0.178	0.218	0.244	0.249	0.265	0.257	0.316
18	7	0.156	0.210	0.239	0.244	0.258	0.257	0.200

Table 2: Performance measures for 1st baseball data set

	$\widehat{TSE}_R^*$	SE	Brier (×10)	SE
MLE	1.000	0.000	2.027	0.029
ARMS	0.343	0.129	2.002	0.027
Stein 1	0.149	0.391	1.993	0.027
Stein 2	0.147	0.416	1.992	0.026
NPEB	0.284	0.277	1.998	0.029
NPML	0.140	0.234	1.991	0.027
Mean	0.193	0.255	1.992	0.025

another one. However, there is a difference between the shrinkage methods and the individual MLEs: for example, the ARMS prediction scores are less than the individual MLE scores in all but two cases.

None of the shrinkage methods do better than using the overall mean. This suggests that more information may need to be gathered in order to distinguish batting performance with confidence.

This example reveals two interesting aspects in relation to ARMS. Firstly, there is little difference between ARMS and NPML predictions, despite the NPML estimate for the random-effects distribution being quite different to the empirical distribution of individual fits used by ARMS. The NPML was estimated by an EM algorithm [18] to have just two atoms at 0.254 and 0.311 with masses 0.797 and 0.203, and log-likelihood -468.675. Secondly, it can be seen that the more sophisticated shrinkage methods are all approximations in practice: Stein estimates are justified because the transformed data will approximate a normal distribution; the NPML only has two mass points and so also clearly approximates the true random-effects distribution.

4.4. Batting averages in 2005

In the previous section the same, moderately large, number of observations is recorded on each of the players prior to prediction. The case where the n_i vary is considered in detail by [6] using data from the 2005 USA baseball season. We use the same data to consider predictions for the second half of the season.

Figure 1 shows a comparison between the Stein estimates of [6], NPML and ARMS. The chart shows that all methods shrink the predictions from the diagonal line of no shrinkage, to the horizontal line of complete shrinkage to the sample mean. The Stein estimates do not take account of the number of attempts to hit a ball because the predictions are shrunk by a constant factor; and of the three shrinkage methods the NPML predictions are shrunk the most, ARMS

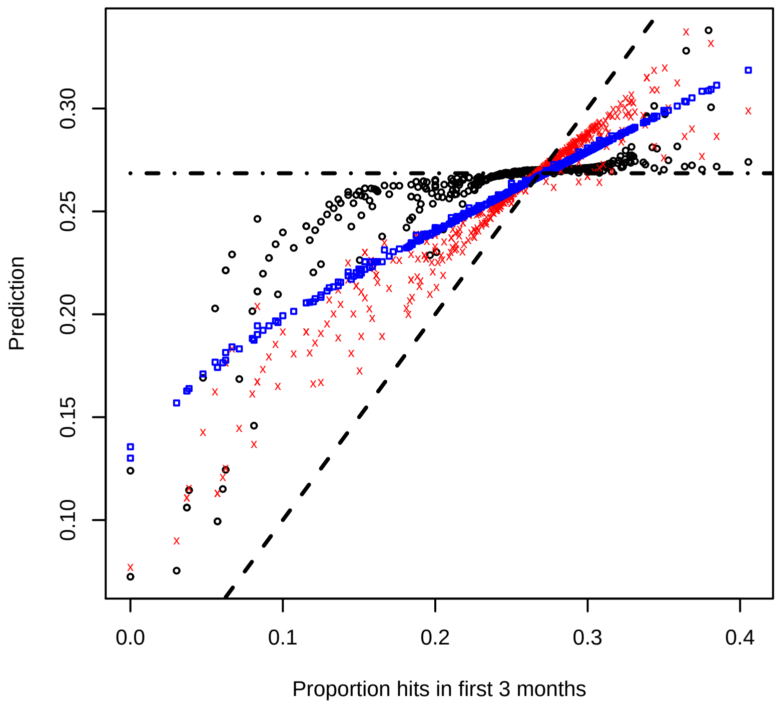

Figure 1: Shrinkage from Stein 2 (square), NPML (o) and ARMS (×) predictions. The diagonal line (– – –) corresponds to predictions as the first-three month individual average, the horizontal line (– . –) to using the overall mean (0.266).

the least. The NPML estimates arise from a discrete distribution estimated using an EM algorithm to have just 4 points $(0.0618, 0.2054, 0.2703, 0.3408)$ with masses $(0.0268, 0.0544, 0.9025, 0.0163)$ and log-likelihood -44203.96.

The performance measure results are in Table 3. They show that the ARMS predictions are not beaten by others. ARMS also performs favourably in comparison with some other methods in Table 2 of [6]. Overall, this example shows that ARMS may again be competitive with more sophisticated methods.

The Stein 1 predictions [13] had a lower $\widehat{TSE}^*_R$ than Stein 2 [6]. This ap-

Table 3: Performance measures for 2nd baseball data set, half-season predictions. Two overall mean definitions are used: the first (0.266) is the number of hits divided by the number of at-bats; the second (0.240) is the average of the MLEs. The second definition was used by [6].

	$\widehat{TSE}_R^*$	SE	Brier (×10)	SE
MLE	1.000	0.000	1.958	0.008
ARMS	0.548	0.086	1.951	0.007
Stein 1	0.484	0.067	1.950	0.008
Stein 2	0.540	0.081	1.949	0.007
NPEB	0.510	0.094	1.949	0.007
NPML	0.717	0.133	1.950	0.007
Mean 1	1.142	0.260	1.951	0.008
Mean 2	0.888	0.233	1.958	0.009

pears to arise because the variance-stabilised variables in [6], and replicated here in Stein 2 and NPEB, were re-transformed to $\hat{u}_i = \sin(\hat{x}_i)^2$, being justified from asymptotic considerations. However, if they are instead re-transformed using the actual transformation, *i.e.* $\hat{u}_i = \{\sin(\hat{x}_i)^2(n_i + 1/2) - 1/4\}/n_i$, then the apparent gain in performance of Stein 1 over Stein 2 disappears and both performance measures decrease. In particular $\widehat{TSE}_R^*$ was improved by 0.056 for the Stein estimates, and by 0.045 for NPEB.

5. Spatial-temporal precipitation prediction

5.1. Introduction

The baseball examples show that ARMS may produce predictions that are competitive with more sophisticated methods in a univariate setting. We stressed the computational advantages of ARMS over some other methods in Section 3, but, of course, Stein estimates are more computationally efficient than ARMS because they do not involve computation of the z_{ij} matrix. The ARMS approach is more useful when analytical results are harder to come by, such as might be the case for more complex models for individual behaviour. One example is found in [2], who used ARMS for predictions on the temporal pattern of individual withdrawals at automated teller machines (ATMs), taking the model

$p(\boldsymbol{y}_i|\boldsymbol{u}_i)$ to be a self-exciting point process with 7 parameters. In other work, the withdrawal amounts were modelled using a multinomial model, and another one that incorporated a form of serial dependence [3]. For the multinomial model ARMS was compared against the use of a Dirichlet distribution for $p(\boldsymbol{u};\boldsymbol{\theta})$, a model that falls into the category described in Section 3.2.

The approach may also be applied when covariates $\boldsymbol{a}_i$ are available for each individual, and conditional predictions are required. We assume that $\boldsymbol{u}_i$ is independent of $\boldsymbol{a}_i$, that is $P_e(\boldsymbol{u}_i = \hat{\boldsymbol{u}}_j|\boldsymbol{a}_i) = P_e(\boldsymbol{u}_i = \hat{\boldsymbol{u}}_j) = 1/n$. Then we re-define z_{ij} as

$$\begin{aligned} z_{ij} &= P_e(\boldsymbol{u}_i = \hat{\boldsymbol{u}}_j|\boldsymbol{y}_i, \boldsymbol{a}_i) \\ &= \frac{p_e(\boldsymbol{y}_i|\boldsymbol{a}_i, \hat{\boldsymbol{u}}_j)}{\sum_{k=1}^{n} p_e(\boldsymbol{y}_i|\boldsymbol{a}_i, \hat{\boldsymbol{u}}_k)}. \end{aligned}$$

The resulting predictions on random variables or events C of interest are

$$p_e(C|\boldsymbol{a}_i, \boldsymbol{y}_i) = \sum_{j=1}^{n} z_{ij} p(C|\boldsymbol{a}_i, \hat{\boldsymbol{u}}_j).$$

We use this approach next to improve weather forecasts.

5.2. Data and methods

The data consist of $n = 444$ individual precipitation-monitoring stations, located over the North American Pacific Northwest. Measurements y_{ij} were made at each station i over a two-year period of $n_i = 686$ days in 2003-05. The data used are available from `http://www.stat.washington.edu/MURI/`, and some of the methods previously applied to the data are in a computer package called `ensembleBMA` [14] for the statistical software R [27]. We review the approach to prediction with these data taken by [31], in the context of ARMS.

Some 56% of the observations are missing, but they are treated as missing at random. There are also forecasts a_{ijk} of precipitation made 48 hours in advance from $k = 1, \ldots, 9$ separate computer models. [31] turned these deterministic predictions into probability densities by using a model for $p(y_{ij}|a_{ijk}, \boldsymbol{u}_{ijk})$ with $\boldsymbol{u}_{ijk} = \boldsymbol{u}_{jk}$ for $i = 1, \ldots, n$. That is, the forecasts of y_{ij} are covariates in a model where the same model parameters are taken for all stations. The $\boldsymbol{u}_{jk}$ have dimension seven: three parameters are used in a logistic regression to model the probability of zero precipitation, and four are used through a gamma distribution to model the precipitation amount when it is not zero.

Parameters $\boldsymbol{u}_{jk}$ for $j = 31, \ldots, 686$ and $k = 1, \ldots, 9$ were fitted in [31] by using the observations y_{ij} and a_{ijk} for $i = 1, \ldots, 444$ from a sliding window of $L = 30$ previous days $\{(j - L), \ldots, (j - 1)\}$. We use the notation $\boldsymbol{y}_{ij}^{(L)} = \{y_{i(j-L)} \ldots, y_{i(j-1)}\}$, and $\boldsymbol{y}_{j}^{(L)} = \{\boldsymbol{y}_{1j}^{(L)}, \ldots, \boldsymbol{y}_{nj}^{(L)}\}$. Similarly, $\boldsymbol{a}_{ijk}^{(L)} = \{a_{i(j-L)k}, \ldots, a_{i(j-1)k}\}$, and a superset is defined when a subscript is dropped, *e.g.* $\boldsymbol{a}_{jk}^{(L)} = \{\boldsymbol{a}_{1jk}^{(L)}, \ldots, \boldsymbol{a}_{njk}^{(L)}\}$, $\boldsymbol{a}_{j}^{(L)} = \{\boldsymbol{a}_{j1}^{(L)}, \ldots, \boldsymbol{a}_{j9}^{(L)}\}$.

The approach in [31] averages spatial heterogeneity in estimation, since $\hat{\boldsymbol{u}}_{ijk}$ is taken to be $\hat{\boldsymbol{u}}_{jk}$ for all i. They used an EM algorithm to estimate $P\{\boldsymbol{u_j} = \hat{\boldsymbol{u}}_{jk}|\boldsymbol{y}_j^{(L)}, \boldsymbol{a}_j^{(L)}; \boldsymbol{\theta}\} = \theta_{jk}$, so that $\sum_{k=1}^{9} \theta_{jk} = 1$. The $\hat{\boldsymbol{\theta}}$ are used to predict precipitation y_{ij} at each site i and time point j through

$$p\{y_{ij}|\boldsymbol{a}_{ij}, \boldsymbol{a}_j^{(L)}, \boldsymbol{y}_j^{(L)}; \hat{\boldsymbol{\theta}}\} \quad = \quad \sum_{k=1}^{9} \hat{\theta}_{jk} p(y_{ij}|a_{ijk}, \hat{\boldsymbol{u}}_{jk}). \tag{6}$$

The basic ARMS method does not average spatial differences across i, and uses

$$\begin{aligned} z_{ijk} &= P_e\{\boldsymbol{u}_{ij} = \hat{\boldsymbol{u}}_{ijk}|\boldsymbol{y}_{ij}^{(L)}, \boldsymbol{a}_{ij}^{(L)}\} \\ &= \frac{p_e\{\boldsymbol{y}_{ij}^{(L)}|\boldsymbol{a}_{ijk}^{(L)}, \hat{\boldsymbol{u}}_{ijk}\}}{\sum_{l=1}^{9} p_e\{\boldsymbol{y}_{ij}^{(L)}|\boldsymbol{a}_{ijl}^{(L)}, \hat{\boldsymbol{u}}_{ijl}\}}, \end{aligned} \tag{7}$$

so that $\sum_{k=1}^{9} z_{ijk} = 1$ for given i and j. The ARMS prediction that is equivalent to (6) replaces $\hat{\theta}_{jk}$ by averaging z_{ijk} over i, that is by using

$$\begin{aligned} z_{.jk} &= P_e\{\boldsymbol{u}_j = \hat{\boldsymbol{u}}_{jk}|\boldsymbol{a}_j^{(L)}, \boldsymbol{y}_j^{(L)}\} \\ &= 1/444\sum_{i=1}^{444} z_{ijk}, \end{aligned}$$

and so $\sum_{k=1}^{9} z_{.jk} = 1$, for given j.

Spatial effects will be present in the 'signal' $\boldsymbol{a}_{ij}$, but there might also be spatial differences in the performance of the weather forecasts across the recording stations. For example, for each time point j the computer model forecast a_{1j1} might be best for station 1, but forecast a_{2j9} best for station 2. Although such potential spatial differences are averaged in estimation, they might be used in prediction. One way to do this is to use

$$\begin{aligned} z'_{ijk} &= P\{\boldsymbol{u}_{ij} = \hat{\boldsymbol{u}}_{ijk}|\boldsymbol{y}_{ij}^{(L)}, \boldsymbol{a}_{ij}^{(L)}; \hat{\boldsymbol{\theta}}\} \\ &= \frac{p\{\boldsymbol{y}_{ij}^{(L)}|\boldsymbol{a}_{ijk}^{(L)}, \hat{\boldsymbol{u}}_{ijk}\}\hat{\theta}_{jk}}{\sum_{l=1}^{9} p\{\boldsymbol{y}_{ij}^{(L)}|\boldsymbol{a}_{ijl}^{(L)}, \hat{\boldsymbol{u}}_{ijl}\}\hat{\theta}_{jl}}. \end{aligned} \tag{8}$$

to make predictions of the form:

$$p\{y_{ij}|\boldsymbol{a}_{ij}, \boldsymbol{a}_{ij}^{(L)}, \boldsymbol{y}_i^{(L)}; \hat{\boldsymbol{\theta}}\} = \sum_{k=1}^{9} z'_{ijk} p\{y_{ij}|a_{ijk}, \hat{\boldsymbol{u}}_{ijk}\}. \tag{9}$$

The ARMS version of prediction equation (9) is to use (7) in place of (8). The difference between these definitions is that the random-effects distribution $\pi(\boldsymbol{u})$ in (8) is taken to put mass θ_{jk} on each point $\hat{\boldsymbol{u}}_{jk}$, rather than 1/9 as in (7). In the remainder of this section we compare the performance of using spatial versus averaged predictions, with the two choices for $\pi(\boldsymbol{u})$. The approach is called Bayesian model averaging (BMA) in [31]. We call it xBMA to denote that the predictions do not take into account possible spatial differences; the ARMS equivalent is denoted xARMS. sBMA and sARMS are used for the predictions

Table 4: Weights (%) used for the chance (%) of rain at two monitoring stations: (a) KCLM on 19 May 2003, (b) KPWT on 26 Jan 2003. It rained at both, with 2 and 26 units recorded respectively. The predictive variance is a plug-in one.

Forecast		(a)				(b)		
Model k	xBMA	xARMS	sBMA	sARMS	xBMA	xARMS	sBMA	sARMS
AVN/GFS	39	19	35	14	30	21	95	88
CENT	0	10	0	13	0	9	0	3
CMCG	0	8	0	5	23	15	0	1
ETA	29	18	34	18	23	13	5	6
GASP	18	11	10	9	0	8	0	0
JMA	14	11	21	24	0	9	0	2
NGPS	0	7	0	14	3	11	0	0
TCWB	0	8	0	1	16	7	0	0
UKMO	0	8	0	2	5	7	0	0
Prediction	17	19	18	19	63	60	64	63
Variance	14	15	14	15	23	23	23	23

Table 5: Average Brier scores (×10) for the four shrinkage predictions, and the best (UKMO) and worst (TCWB) performing individual forecasts. Predictions and outcomes are for whether it will rain, over the period from 11th December 2002 to 31st March 2005.

xBMA	xARMS	sBMA	sARMS	UKMO	TCWB
1.410	1.414	1.427	1.427	1.470	1.577

from equation (9).

In this example xARMS does not really offer a substantial saving in computational cost over xBMA. The example is better viewed as a controlled attempt to compare any loss in performance from using the approximate approach to obtain weights in model averaging, rather than optimisation. That is, the only difference between the ARMS and BMA predictions that follow is the weights: the model parameters are the same.

5.3. Results

Predictions for each method are formed from weighted sums. Table 4 shows the different weights assigned to the 9 computer model forecasts for two monitoring stations and time points. A similar table is found in [31], and further details of the computer forecasts are given therein. The xBMA and xARMS pre-

Table 6: Percentage of precipitation monitoring stations for which the method labelled in the first column has a lower average Brier score that the method in the top row. Predictions are for whether it will rain at each station. The best (UKMO) and worst (TCWB) performing individual forecasts under the overall Brier score are included for comparison.

	xBMA	xARMS	sBMA	sARMS	UKMO	TCWB
xBMA	0	62	74	71	86	96
xARMS	38	0	65	72	84	96
sBMA	26	35	0	53	75	95
sARMS	29	28	47	0	74	96
UKMO	14	16	25	26	0	86
TCWB	4	4	5	4	14	0

dictions apply one set of weights to all monitoring stations at each time point, but the weights vary across stations for sBMA and sARMS. The table shows that there is a difference between the weights. For example, in column (b) the AVN/GFS forecast receives a much higher weighting under sBMA and sARMS. However, the predictions shown at the bottom of the table are broadly similar.

To assess the performance across the monitoring sites and predictions made through time we focus on assessing predictions about the probability of rain, using average Brier scores. Table 5 shows that xBMA and xARMS perform better than sBMA and sARMS, but that all do better than the top-performing single density prediction based on the United Kingdom Met Office (UKMO) forecasts. It is tricky to obtain bootstrap estimates of standard errors of the Brier scores because of the temporal nature of the data. However, the difference between the methods may be further investigated by examining the performance for the $i = 1, \ldots, 444$ monitoring stations in Table 6. Reading along a row shows the percentage of monitoring stations in which the row label beat the column label under an average Brier score. It shows that no method outperformed the others all the time: even the worst performing overall TCWB forecasts did better than xBMA for some monitoring stations. The table also suggests that, for a given time point, performance is worse when the weights in the predictions are allowed to vary across the stations.

For an overall Brier score performance measure, it is not worth using ARMS to account for spatial effects beyond the information in the $\boldsymbol{a}_{ij}$ terms. This might also be the case for more complicated models. A Gaussian model for spatial variability additional to that in the density predictions based on the UKMO deterministic forecasts was built by [1]. The model was evaluated in a similar manner to Table 5 and had a Brier score of 0.148, which is not an improvement over using the UKMO density predictions alone.

5.4. Remarks

The use of ARMS in this section shows a relatively simple way to investigate whether improvements may be gained through modelling spatial effects beyond the ‘signal’ $\boldsymbol{a}_{ij}$ from deterministic computer forecasts. Under a Brier score for the data analysed, we found that it is not worth modelling additional spatial variability. This was also found in some other work using the same data set, but a more sophisticated model [1]. The result is likely to be because there is very little spatial heterogeneity in relation to the noise in the system. Applying such a model leads to a worsening in performance because the estimates for an individual monitoring site are better shrunk all the way to the mean.

The example also showed that predictions in which the random-effects distribution $\pi(\boldsymbol{u})$ is taken to put mass θ_{jk} on each point $\hat{\boldsymbol{u}}_{jk}$ might perform slightly better than using equal weights 1/9. The latter approach has the benefit of not requiring optimisation. But, in the context of the general problem described in Section 1, fitting individual parameters and then estimating weights is an interesting extension to be investigated in future work.

6. Conclusion

In this article we have formalised a general shrinkage approach (ARMS) for repeated-measures data. Using some of the terminology from relevant previous work, the method might be described as approximate non-parametric empirical Bayes prediction. The main advantage of ARMS is that it is computationally feasible to apply to a range of problems, and that its shrinkage estimates might not practically differ from those obtained by more sophisticated methods. Two examples were presented to investigate the legitimacy of this claim. In the first some baseball data with univariate random-effect binomial models were re-examined. The method was found to perform favourably in comparison with Stein and other empirical Bayes predictions. In the second example ARMS was applied to a spatial-temporal forecasting problem with multivariate random-effects, and was shown to provide a relatively easy way to check whether performance might be improved by modelling spatial effects.

In a famous piece of statistical folk lore, the Box-Jenkins forecasting procedure was variously compared to a Rolls Royce, or a racing car [9]. Following the automobile analogy, we do not claim that ARMS is the leanest, meanest Formula 1 car around: more sophisticated methods might do better in certain situations. However, it will usually have better predictive performance than the family hatchback of modelling individuals separately. ARMS is perhaps most akin to a Morris Mini Cooper S: it is a versatile vehicle that can take a few knocks in a range of environments without too much specialist training, and it will get you there relatively quickly.

Acknowledgments

Adam Brentnall's work was supported by EPSRC grant number EP/D505380/1 and the work of David Hand was partially supported by a Royal Society Wolfson Research Merit Award.

References

[1] Berrocal, V. J., Raftery, A. E. and Gneiting T., 2008. Probabilistic quantitative precipitation forecasting using a two-stage spatial model. Ann. Appl. Statist., 2, 1170–1193.

[2] Brentnall, A. R., Crowder, M. J. and Hand D. J., 2008. A statistical model for the temporal pattern of individual automated teller machine withdrawals. Appl. Statist., 57, 43–59.

[3] Brentnall, A. R., Crowder, M. J. and Hand, D. J., 2010. Predicting the amounts that individuals withdraw at cash machines using a random effects multinomial model. Statistical Modelling, 10, 197–214.

[4] Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1–3.

[5] Brown, L. D., 1971. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist., 42, 855–903.

[6] Brown, L. D., 2008. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. Ann. Appl. Statist., 2, 113–152.

[7] Brown, L. D. and Greenshtein, E., 2009. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. Ann. Statist., 37, 1685–1704.

[8] Caffo, B. S., Jank, W. and Jones, G. L., 2005. Ascent-based Monte Carlo expectation- maximization. J. R. Statist. Soc. B, 67, 235–251.

[9] Chatfield, C. and Prothero, D. L., 1973. Box-Jenkins seasonal forecasting: Problems in a case-study (with discussion). J. R. Statist. Soc. A, 136, 295–336.

[10] Copas, J. B., 1969. Compound decisions and empirical Bayes. J. R. Statist. Soc. B, 31, 397–425.

[11] Efron, B., 2003. Robbins, empirical Bayes and microarrays. Ann. Statist., 31, 366–378.

[12] Efron, B. and Morris, C., 1973. Combining possibly related estimation problems. J. R. Statist. Soc. B, 35, 379–421.

[13] Efron, B. and Morris, C., 1975. Data analysis using Stein's estimator and its generalizations. J. Am. Statist. Ass., 70, 311–319.

[14] Fraley, C., Raftery, A. E., Gneiting, T. and Sloughter, J. M., 2008. EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging. Technical Report 516R, Department of Statistics, University of Washington.

[15] Gao, J. and Hitchcock, D. B., 2010. James-Stein shrinkage to improve k-means cluster analysis. Comput. Statist. Data Anal., 54, 2113–2127.

[16] Gneiting, T. and Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Statist. Ass., 102, 359–378.

[17] James, W. and Stein, J., 1961. Estimation with quadratic loss. In: J. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif., University of California Press, 1, 361–379.

[18] Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Statist. Ass., 73, 805–811.

[19] Laird, N. M. and Louis, T. A., 1991. Smoothing the non-parametric estimate of a prior distribution by roughening : A computational study. Comput. Statist. Data Anal., 12, 27–37.

[20] Lesperance, M. L. and Kalbfleisch J. D., 1992. An algorithm for computing the nonparametric MLE of a mixing distribution. J. Am. Statist. Ass., 87, 120–126.

[21] An, L., Nkurunziza, S., Fung, K. Y., Krewski, D. and Luginaah, I., 2009. Shrinkage estimation in general linear models. Comput. Statist. Data Anal., 53, 2537–2549.

[22] Lindley, D. V. and Smith, A. F. M., 1972. Bayes estimates for the linear model. J. R. Statist. Soc. B, 34, 1–41.

[23] MacEachern, S. N. and Müller, P., 1998. Estimating mixture of Dirichlet process models. J. Comput. Graph. Statist., 7, 223–238.

[24] McAuliffe, J., Blei, D. and Jordan, M., 2006. Nonparametric empirical Bayes for the Dirichlet process mixture model. Statist. Comput., 16, 5–14.

[25] Morris, C. N., 1983. Parametric empirical Bayes inference: Theory and applications. J. Am. Statist. Ass., 78, 47–55.

[26] Müller, P. and Quintana, F. A., 2004. Nonparametric Bayesian data analysis. Statist. Sci., 19, 95–110.

[27] R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[28] Robbins, H., 1951. Asymptotically subminimax solutions of compound statistical decision problems. In: J. Neyman (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif., University of California Press, 131–148.

[29] Robbins, H., 1956. An empirical Bayes approach to statistics. In: J. Neyman (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical

Statistics and Probability Berkeley, Calif., University of California Press, 1, 157–163.

[30] Robbins, H., 1964. The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1–20.

[31] Sloughter, M. J., Raftery, A. E., Gneiting, T. and Fraley C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly Weather Review, 135, 3209–3220.

[32] Stein, J., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: J. Neyman (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif., University of California Press, 1, 197–206.

[33] Stigler, S. M., 1990. The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. Statist. Sci., 5, 147–155.

[34] Wang, Y., 2007. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J. R. Statist. Soc. B, 69, 185–198.