SUPPLEMENT TO "STATISTICAL UNFOLDING OF ELEMENTARY PARTICLE SPECTRA: EMPIRICAL BAYES ESTIMATION AND BIAS-CORRECTED UNCERTAINTY QUANTIFICATION"

By Mikael Kuusela and Victor M. Panaretos

École Polytechnique Fédérale de Lausanne

This online supplement complements the main paper by providing additional simulation results as well as some technical details. In Section 1, we compare the empirical Bayes and the hierarchical Bayes approaches to unfolding. Section 2 contains technical material on the convergence and mixing of the single-component Metropolis–Hastings sampler and on the Gaussian approximation used in the coverage studies, while Section 3 provides convergence studies for empirical Bayes estimation as well as a detailed comparison of the various types of confidence intervals considered in this work.

1. Comparison of empirical Bayes and hierarchical Bayes.

1.1. *Hierarchical Bayes as an alternative to empirical Bayes.* The fully Bayesian alternative to empirical Bayes for handling the unknown hyperparameter δ is to consider a Bayesian hierarchical model where δ is given a hyperprior $p(\delta)$. This allows one to form the joint posterior of $\boldsymbol{\beta}$ and δ,

$$p(\boldsymbol{\beta}, \delta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\delta)p(\delta)}{p(\boldsymbol{y})}, \tag{1}$$

after which $\boldsymbol{\beta}$ can be estimated using the mean of the marginal posterior

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \int_{\mathbb{R}_+} p(\boldsymbol{\beta}, \delta|\boldsymbol{y})\,\mathrm{d}\delta. \tag{2}$$

We consider prior distributions of the form

$$p(\delta) \propto \mathbf{1}_{[L,\infty)}(\delta)\,\delta^{a-1}e^{-b\delta}, \tag{3}$$

with the parameters a, b and L chosen in such a way that the density can be normalized. This family of priors includes as special cases the Pareto$(-a, L)$ distribution, which is obtained by taking $a < 0$, $b = 0$ and $L > 0$; and the Gamma(a, b) distribution obtained with $a > 0$, $b > 0$ and $L = 0$. The full

posterior conditional for δ is given by

$$p(\delta|\boldsymbol{\beta}, \boldsymbol{y}) = p(\delta|\boldsymbol{\beta}) \propto p(\boldsymbol{\beta}|\delta)p(\delta) \tag{4}$$

$$\propto \mathbf{1}_{[L,\infty)}(\delta)\, \delta^{p/2+a-1} \exp\left(-(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta} + b)\delta\right), \tag{5}$$

which shows that the prior family (3) is conditionally conjugate. In particular, $p(\delta|\boldsymbol{\beta}, \boldsymbol{y})$ is the Gamma$(p/2 + a, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta} + b)$ distribution truncated to the interval $[L, \infty)$ provided that $p/2 + a > 0$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta} + b > 0$. As a result, sampling from the full posterior $p(\boldsymbol{\beta}, \delta|\boldsymbol{y})$ can be implemented as a simple extension of the Gibbs sampler underlying the single-component Metropolis–Hastings algorithm described in Section 4.3.1 of the main paper: we loop over all the unknowns and for each β_k we sample from the corresponding approximate full posterior conditional using a Metropolis–Hastings acceptance step to correct for the approximation, while for δ we simply sample from (5) without a Metropolis–Hastings correction.

Although the hierarchical Bayes model is attractive from the purely Bayesian perspective, it suffers from the problem of requiring the specification of the hyperprior $p(\delta)$. Following purely Bayesian thinking, the hyperprior should be chosen based on the analyst's subjective degree of belief regarding the value of δ. But, in a typical high energy physics experiment involving thousands of scientists, a consensus on a specific hyperprior is unlikely, especially for such an abstract quantity as the regularization parameter δ. Hence the only reasonable choice would be an uninformative flat prior, but this requires the specification of the metric in which $p(\delta)$ is flat. Unfortunately, hierarchical Bayes estimation is known to be sensitive to this non-trivial choice (Gelman, 2006), which is also what we observe in our simulations in Sections 1.2 and 3.2 below. In these simulations, empirical Bayes on the other hand achieves performance which is comparable with uninformative hierarchical Bayes without the need to make any extra distributional assumptions on δ. By construction, the method is invariant to transformations of δ and free of tuning parameters once the family of regularizing priors has been selected. As empirical Bayes allows the data analyst to achieve good performance without the need for hyperprior elicitation and sensitivity analysis, we feel that it provides a more appealing solution to the HEP unfolding problem than hierarchical Bayes.

1.2. *Simulation study.* In this section, we compare the point estimation performance of empirical Bayes and hierarchical Bayes using the simulation setup of Section 5 of the main paper. The performance of the methods is compared using the integrated squared error ISE $= \int_E (\hat{f}(s) - f(s))^2 \,\mathrm{d}s$. We consider the following uninformative, yet proper hyperpriors for δ:

(a) Pareto$(1, 10^{-10})$, (b) Pareto$(1/2, 10^{-10})$, (c) Gamma$(0.001, 0.001)$ and (d) Gamma$(1, 0.001)$. The reasoning behind these choices is that each of these hyperpriors is nearly uniform for some transformation of δ. More specifically, prior (a) is nearly flat for $1/\delta$, (b) for $1/\sqrt{\delta}$, (c) for $\log(\delta)$ and (d) for δ. Out of these, hyperprior (c) is extensively used in the literature (see, e.g., Browne and Draper (2006); Ruppert, Wand and Carroll (2003), Section 16.3; or Young and Smith (2005), Section 3.8), but Gelman (2006) argues that (b) should provide better estimates. Hyperprior (a) is considered by, e.g., Browne and Draper (2006), while hyperprior (d) is flat for the untransformed hyperparameter δ itself. The starting point of the MCMC sampler was $(\boldsymbol{\beta}_{\text{init}}, \delta^{(0)})$ and the rest of the parameters were set to the same values as in the empirical Bayes simulations reported in Section 5 of the main paper. The performance of the methods is compared using 1 000 repeated observations from the smeared process N.

Figure 1 shows boxplots of the relative pairwise ISE differences, $(\text{ISE}_{\text{HB},i} - \text{ISE}_{\text{EB},i})/\text{ISE}_{\text{EB},i}$, between empirical Bayes (EB) and the alternative hierarchical Bayes (HB) models. All the differences between EB and HB are statistically significant at any reasonable significance level, except for the case of hyperprior (c) at $\lambda_{\text{tot}} = 10\,000$, which is only significant at the 1 % level (two-sided Wilcoxon signed-rank test p-value 0.0052), and the same comparison at $\lambda_{\text{tot}} = 20\,000$, which is not statistically significant (two-sided Wilcoxon signed-rank test p-value 0.95).

Two important conclusions emerge from these results. Firstly, there are marked differences in the performance of hierarchical Bayes between the different hyperpriors, especially when there is only a limited amount of data. For example, in the case of the small sample size, the median performance of hierarchical Bayes ranges from 17 % better to 30 % worse than that of empirical Bayes. These differences can be explained in terms of how strongly each hyperprior tends to regularize, with priors that favor small values of δ generally performing better, and vice versa. In particular, hyperprior (d) places too much importance on large hyperparameter values and hence regularizes too strongly. This shows that in the present problem the choice of the hyperprior has an undesirably large influence on the performance of the hierarchical Bayes estimates, unless there is a large amount of data available. Secondly, the superiority of empirical Bayes and hierarchical Bayes depends on which hyperprior is used. Generally, the performance of empirical Bayes is very similar to that of hyperprior (c), with priors (a) and (b) yielding better and prior (d) worse point estimation performance. Unfortunately, there are no guarantees that hyperprior (a) would always perform the best for all types of true intensities. What we can conclude, however, is

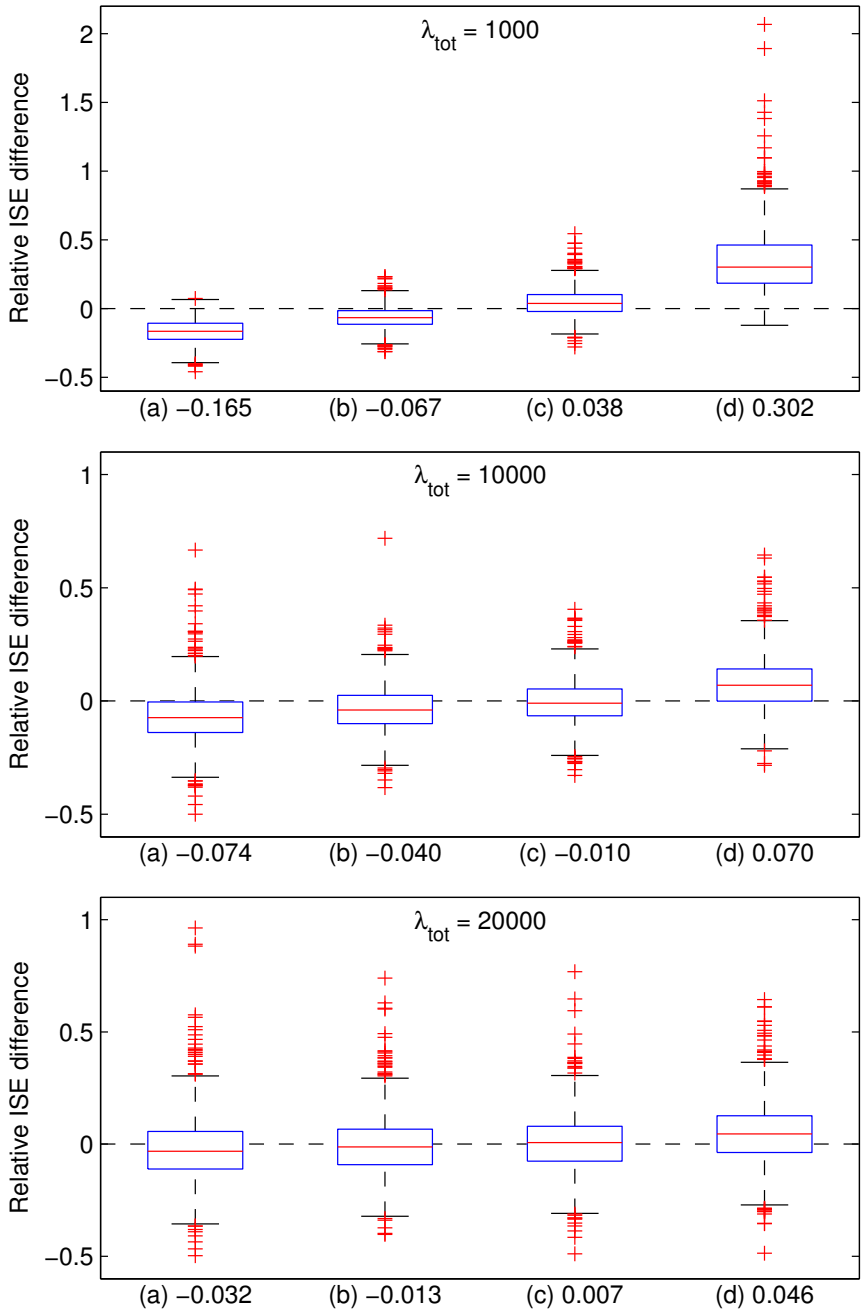

FIG 1. *Boxplots of relative pairwise integrated squared error differences between empirical Bayes and hierarchical Bayes for the following uninformative hyper-priors: (a)* Pareto$(1, 10^{-10})$, *(b)* Pareto$(1/2, 10^{-10})$, *(c)* Gamma$(0.001, 0.001)$ *and (d)* Gamma$(1, 0.001)$. *The numbers below the plots show the medians of the relative differences. A positive difference indicates that hierarchical Bayes incurred a larger error than empirical Bayes.*

that the performance of empirical Bayes is by all means comparable to that of hierarchical Bayes and it achieves this *without making any extra assumptions about the distribution of* δ. In other words, empirical Bayes achieves comparable performance while making only the bare minimum number of assumptions about the unknown intensity.

2. Technical details.

2.1. *Convergence and mixing of the MCMC sampler.* During each iteration of the Monte Carlo expectation-maximization algorithm used in the empirical Bayes estimation of δ, we verify the convergence and mixing of the MCMC sampler by monitoring the acceptance rates of the Metropolis–Hastings proposals and the autocorrelation times κ_j, $j = 1, \ldots, p$, of the Markov chain. The latter measure how often the sampler on average produces an independent observation from the posterior and is estimated using Geyer's initial convex sequence estimator (ICSE) (Geyer, 1992) computed using the R package `mcmc` (Geyer and Johnson, 2013). The autocorrelation times κ_j enable us to define the effective sample sizes $\mathrm{ESS}_j = S/\kappa_j$, $j = 1, \ldots, p$, where S is the size of the MCMC sample. ESS_j measures the effective number of independent observations obtained for the jth component of the Markov chain (Kass et al., 1998, p. 99). For the MCMC run producing the final point estimate $\hat{\boldsymbol{\beta}}$, we also monitor the trace plots, histograms, estimated autocorrelation functions and cumulative means of each component β_j, $j = 1, \ldots, p$, of the Markov chain. In the case of the hierarchical Bayes model, similar diagnostics were also produced for the hyperparameter δ.

Let us first consider the performance of the MCMC sampler in empirical Bayes estimation in the Gaussian mixture model experiments of Section 5 of the main paper. In the case of the large sample size, the autocorrelation time of the MCMC sampler averaged over the components of $\boldsymbol{\beta}$ varied between 4.0 and 9.3 during the convergence of the MCEM iteration. A typical proposal acceptance rate[1] was 98 %. For the final MCMC run producing the point estimate $\hat{\boldsymbol{\beta}}$, a more careful performance analysis was carried out for each component of the sampler using the diagnostic plots described above. These plots indicated no major issues with the convergence and mixing of the sampler. Figure 2 shows these diagnostic plots for the components β_5 and β_{21} after the removal of burn-in. These plots indicate that the chain has converged and mixes reasonably well, although the performance of the chain is typically slightly better in the interior of the space (β_{21}) than closer to

[1] As opposed to the standard multivariate Metropolis–Hastings algorithm, the single-component version of the algorithm tries to imitate the Gibbs sampler and hence the ideal acceptance rate would be 100 %.

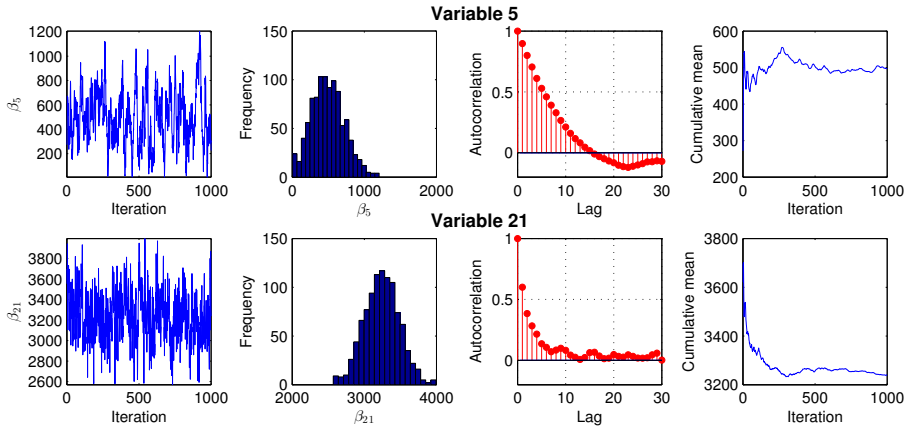

FIG 2. *Convergence and mixing diagnostics for the single-component Metropolis–Hastings sampler for variables β_5 and β_{21}: from left to right, the trace plots, histograms, estimated autocorrelation functions and cumulative means of the samples. For variable β_5, the acceptance rate was 97.6 %, the lag 1 autocorrelation 0.90 and the autocorrelation time 12.2. Hence the effective sample size for β_5 was 81.7. For β_{21}, the corresponding values were 99.5 %, 0.60 and 5.3 with effective sample size of 187.3.*

the boundaries (β_5). Similar, and occasionally even better, sampling performance was also observed in the medium and small sample size experiments and no major issues were identified with the MCMC sampler in these cases either.

Similar checks were also performed for the hierarchical Bayes model studied in Section 1. With the large and medium sample sizes, the MCMC sampler performed just as well as in the case of empirical Bayes estimation in sampling $\boldsymbol{\beta}$. Depending on the sample size and the hyperprior used, the autocorrelation times for the hyperparameter δ varied in the range 5.8–8.8 corresponding to effective sample sizes 113.9–172.2, and the diagnostic plots revealed no problems with the convergence and mixing of δ. But the mixing of the chain for the hierarchical Bayes model with the small sample size leaves some room for improvement. For hyperpriors (a)–(d), the autocorrelation times were 28.1, 18.8, 54.1 and 34.2, respectively. In particular, in the case of hyperprior (c), this corresponds to an effective sample size of only 18.5. Also from the diagnostic plots, it was evident that all these chains were more autocorrelated for δ than desired, although the traceplots still exhibited fairly good exploration of the hyperparameter space. The slow mixing of the hyperparameter also affected to some degree the mixing of the spline coefficients $\boldsymbol{\beta}$, although their autocorrelation times were not as badly inflated as for the hyperparameter.

With such a relatively slowly mixing sampler, one might wonder whether any of our results would change if the chain was let to run for a larger number of time steps. To make sure that this is not the case, we repeated the small sample size simulations in Figure 1 using 5 000 post-burn-in observations from the hierarchical Bayes posterior. As expected, the distributions of the relative ISE differences were more concentrated than previously, but the medians barely changed. Similarly, no major differences were observed in the coverage performance of the corresponding credible intervals studied in Section 3.2.

In the case of the Z boson demonstration, a slight overparameterization of the unfolded space was helpful in facilitating the convergence and mixing of the MCMC sampler, see Section 6.2 of the main paper. With this overparameterization, the sampler performed almost as well as in the Gaussian mixture model experiments and no obvious problems were identified in any of the checks mentioned above. The mixing of the sampler was particularly good around the mode of the Z mass peak, with slightly slower mixing observed in the tails of the intensity.

2.2. *Gaussian approximation for coverage studies.* Estimating the coverage probability of the confidence intervals described in Section 4.4 of the main paper requires running the procedure for many repeated observations which is impractical due to the high computational cost of the MCMC sampler needed for computing the means of the empirical Bayes posterior. We describe in this section a Gaussian approximation to the Poisson likelihood that enables us to compute an approximation to the actual posterior mean in a fraction of the time required for the full MCMC. We then use this Gaussian approximated point estimate $\hat{\boldsymbol{\beta}}_{\mathrm{G}}$ in place of the actual posterior mean $\hat{\boldsymbol{\beta}}$ in the procedure of Section 4.4 to form Gaussian approximated confidence bands for which a coverage study can be carried out. These Gaussian approximated intervals look similar to the full Poisson intervals and we expect the coverage of the full intervals to be similar to or better than what is observed for the Gaussian intervals.

Our data is generated by the model

$$\boldsymbol{y} \sim \mathrm{Poisson}(\boldsymbol{K}\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}_+^p. \tag{6}$$

For large enough Poisson intensities, this can be approximated as

$$\boldsymbol{y} \overset{a}{\sim} N(\boldsymbol{K}\boldsymbol{\beta}, \mathrm{diag}(\boldsymbol{K}\boldsymbol{\beta})), \tag{7}$$

where the covariance can be estimated as $\hat{\boldsymbol{C}} = \mathrm{diag}(\boldsymbol{y}_+)$, where $\boldsymbol{y}_+$ denotes the observed data $\boldsymbol{y}$ with the potential zero counts replaced by ones in order

to make $\hat{\boldsymbol{C}}$ positive definite. The approximate model becomes

$$\boldsymbol{y} \overset{a}{\sim} N(\boldsymbol{K}\boldsymbol{\beta}, \hat{\boldsymbol{C}}), \tag{8}$$

and therefore the empirical Bayes posterior satisfies

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\delta}) \overset{a}{\propto} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta})^{\mathrm{T}}\hat{\boldsymbol{C}}^{-1}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta}) - \hat{\delta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta}\right), \tag{9}$$

where $\overset{a}{\propto}$ denotes 'approximately proportional to'. Ignoring the positivity constraint, this is a multivariate Gaussian posterior whose mean and mode coincide. Therefore the original posterior mean approximately satisfies:

$$\hat{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\delta}) \approx \operatorname*{arg\,max}_{\boldsymbol{\beta}\in\mathbb{R}^p} -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta})^{\mathrm{T}}\hat{\boldsymbol{C}}^{-1}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta}) - \hat{\delta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta} \tag{10}$$

$$= \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta})^{\mathrm{T}}\hat{\boldsymbol{C}}^{-1}(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta}) + 2\hat{\delta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{A}}\boldsymbol{\beta} \tag{11}$$

$$= (\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{C}}^{-1}\boldsymbol{K} + 2\hat{\delta}\boldsymbol{\Omega}_{\mathrm{A}})^{-1}\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{C}}^{-1}\boldsymbol{y} := \hat{\boldsymbol{\beta}}'_{\mathrm{G}}, \tag{12}$$

where we have used the positive definiteness of $\boldsymbol{\Omega}_{\mathrm{A}}$ and the hyperparameter estimate $\hat{\delta}$ is given by the MCEM iteration. As the final step of this procedure, we enforce the positivity constraint of $\boldsymbol{\beta}$ by setting any negative values in $\hat{\boldsymbol{\beta}}'_{\mathrm{G}}$ to zero, that is,

$$\hat{\boldsymbol{\beta}}_{\mathrm{G}} = \mathbf{1}\left\{\hat{\boldsymbol{\beta}}'_{\mathrm{G}} \geq \mathbf{0}\right\} \circ \hat{\boldsymbol{\beta}}'_{\mathrm{G}}, \tag{13}$$

where $\circ$ denotes the elementwise product of two vectors. With these approximations, the main computational cost in forming the estimate $\hat{\boldsymbol{\beta}}_{\mathrm{G}}$ comes from the matrix operations in Equation (12) which are many orders of magnitude faster than running the MCMC sampler.

3. Additional simulation results. In this section, we provide further simulation results to complement those presented in Section 5 of the main paper. Unless otherwise stated, all these results concern the Gaussian mixture model setup described in Section 5.1 of the main paper.

3.1. *Convergence studies.* We verified that 30 iterations were sufficient for the convergence of the MCEM algorithm using Figure 3(a). We see that in the small sample size case, the MCEM iteration increased the regularization strength from the initial value $\delta^{(0)} = 1 \cdot 10^{-5}$, while in the medium and large sample size cases the regularization strength was decreased from its initial value. The algorithm converged the faster the larger the sample size.

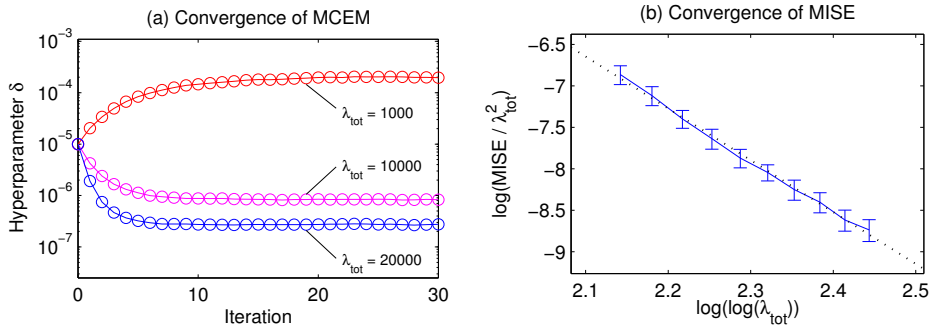

FIG 3. *Convergence studies for empirical Bayes unfolding. Figure (a) illustrates the convergence of the Monte Carlo EM algorithm and shows that the algorithm converges faster for larger sample sizes. Figure (b) shows the convergence of the mean integrated squared error (MISE) as the expected sample size* λ_{tot} *grows. Note that convergence is only obtained for* $\text{MISE}/\lambda_{\text{tot}}^2$. *The error bars indicate approximate 95 % confidence intervals, and the dotted straight line is a least-squares fit to the convergence curve.*

With the small sample size, the algorithm took approximately 23 iterations to converge, while in the medium sample size case this was reduced to 15 iterations and further down to 10 iterations with the large sample size. In each case, there was little Monte Carlo variation in the hyperparameter estimates. A similar plot was used to verify the convergence of the MCEM algorithm in the Z boson experiments of Section 6 of the main paper in which case the algorithm converged in roughly 10 iterations.

To further study how empirical Bayes point estimation behaves as a function of the sample size, we repeated the Gaussian mixture model experiment on a logarithmic grid of expected sample sizes ranging from $\lambda_{\text{tot}} = 5\,000$ up to $\lambda_{\text{tot}} = 100\,000$. For each sample size, we unfolded 100 independent realizations of the smeared data $\boldsymbol{y}$ and estimated the mean integrated squared error (MISE) of $\hat{f}$ as the sample mean of the integrated squared errors $\text{ISE} = \int_E (\hat{f}(s) - f(s))^2 \, \mathrm{d}s$. As $\lambda_{\text{tot}} \to \infty$, we expect the MISE to diverge, but $\text{MISE}/\lambda_{\text{tot}}^2$ should converge towards zero, and this is indeed what we observe in Figure 3(b).

In the classical problem of deconvolving a density function smeared by Gaussian noise, the optimal convergence rate of the MISE is known to be of the order $(\log n)^{-k}$ (Meister, 2009), where n is the number of i.i.d. smeared observations and k depends on the smoothness of the true density. Our setup differs slightly from the classical density deconvolution problem in the sense that we observe a realization of a smeared Poisson point process and try to estimate the intensity function of the corresponding true process. We have also performed all the computations on a compact interval, which introduces

boundary effects near the end points of the interval. Nevertheless, one might conjecture that $\mathrm{MISE}/\lambda_{\mathrm{tot}}^2$ converges at the rate $(\log \lambda_{\mathrm{tot}})^{-k}$ in which case one would expect the values in Figure 3(b) to fall on a straight line with slope $-k$. This indeed seems to approximately be the case: the $\mathrm{MISE}/\lambda_{\mathrm{tot}}^2$ values seem to follow fairly well the line with slope $-k = -6.25$, which is also shown in the figure. However, in a more careful inspection, the convergence curve appears to have a slightly convex shape which possibly indicates that the actual convergence rate is somewhat slower than $(\log \lambda_{\mathrm{tot}})^{-k}$. This might be due to the fact that we kept the number of basis functions fixed while increasing λ_{tot}. As a result, the discretization error from the spline fit should eventually slow down the convergence rate.

3.2. *Detailed comparison of confidence intervals.* The aim of this section is to provide a comprehensive comparison of the various types of confidence intervals considered in the simulation experiments of Section 5 in the main paper. Additionally, we also consider the performance of the credible intervals of the hierarchical Bayes model of Section 1 in this supplement. On the following pages, we provide for each sample size the following figures:

(a) The iteratively bias-corrected confidence intervals for the full Poisson likelihood
(b) The same intervals, but for a Gaussian approximation of the Poisson likelihood
(c) Empirical coverage for the different intervals
(d) Empirical coverage for the iteratively bias-corrected intervals with varying amounts of bias correction
(e) Observed intervals for the methods considered in plot (c)
(f) Observed intervals for the varying amounts of bias correction considered in plot (d)

The figures are presented first for the small sample size, followed by the medium and large sample size cases.

Figure 4 shows a comparison of the iteratively bias-corrected intervals obtained with the full Poisson likelihood and with the Gaussian approximation of Section 2.2 for $\lambda_{\mathrm{tot}} = 1\,000$ and $N_{\mathrm{BC}} = 15$. The corresponding point estimates are also shown in the figure. We observe that the Gaussian approximated intervals are very similar to the full intervals so it is reasonable to expect that their coverage properties would also be similar.

Figure 5(a) compares the empirical coverage of the 95 % iteratively bias-corrected intervals with various forms of Bayesian intervals and standard bootstrap intervals, see Section 5.2.2 of the main paper for a description of the empirical Bayes and the bootstrap intervals. The hierarchical Bayes in-

tervals are the 95 % equal-tailed credible intervals induced by the marginal posterior (2). The bias-corrected intervals and the bootstrap intervals are formed using the Gaussian approximation to the likelihood, while the Bayesian intervals are for the full Poisson likelihood. We observe that both the Bayesian intervals and the standard bootstrap intervals suffer from major undercoverage in areas of sizable bias. On the other hand, the iteratively bias-corrected intervals perform considerably better and achieve close-to-nominal coverage on most parts of the spectrum, except for the peak at $s = 2$ where the empirical coverage drops to 85.0 %.

Figure 5(b) shows the empirical coverage of the Gaussian approximated iteratively bias-corrected intervals for varying amounts of bias correction. The graph for $N_{\mathrm{BC}} = 0$ corresponds to the standard bootstrap percentile intervals. We observe that by increasing the number of bias-correction iterations, we can consistently improve the coverage. In particular, for $N_{\mathrm{BC}} = 50$, the coverage is close to the nominal value across the whole spectrum.

To gain further insight into the coverage results in Figure 5(a), we plot in Figure 6 a single realization of each of the various intervals (for hierarchical Bayes, only the extremal cases corresponding to hyperpriors (a) and (d) are shown). From this figure, it is clear that only the iteratively bias-corrected intervals adequately account for the bias at high curvature regions of the spectrum. The Bayesian intervals are consistently wider than the standard bootstrap intervals, which can be understood in terms of these intervals partially accounting for the bias (Ruppert, Wand and Carroll, 2003, Chapter 6), but evidently some bias still remains unaccounted for. Interestingly, there is little difference in the lengths of the empirical Bayes and hierarchical Bayes intervals. The bootstrap basic intervals can also be seen to partially accommodate the bias, but these intervals are generally too short and the implicit bias correction too small to yield satisfactory coverage.

Figure 7 shows a single realization of the bias-corrected intervals studied in Figure 5(b). We see that as the number of bias correction iterations is increased, the intervals become better centered around the true intensity but at the same time their length is increased. As discussed in Section 4.4 of the main paper, this phenomenon is entirely expected, but the curious part of these results is the fact that even with a large number of bias-correction iterations the interval length is only modestly increased while the coverage performance is significantly improved. It seems that there is a small amount of residual bias that remains after the bias correction and that this residual bias is adequate to regularize the interval length while having only a small effect on the coverage performance.

Figures 8–11 provide the same plots for the medium sample size case with $\lambda_{\text{tot}} = 10\,000$ and $N_{\text{BC}} = 5$, while the large sample size case with $\lambda_{\text{tot}} = 20\,000$ and $N_{\text{BC}} = 5$ is considered in Figures 12–15. Similar conclusions as above emerge from these figures, although the differences become less pronounced as the sample size increases. In each case, the coverage performance of the iteratively bias-corrected intervals is better than that of the competing methods and this difference can be understood in terms of how the different types of intervals account for the bias of the underlying point estimate.

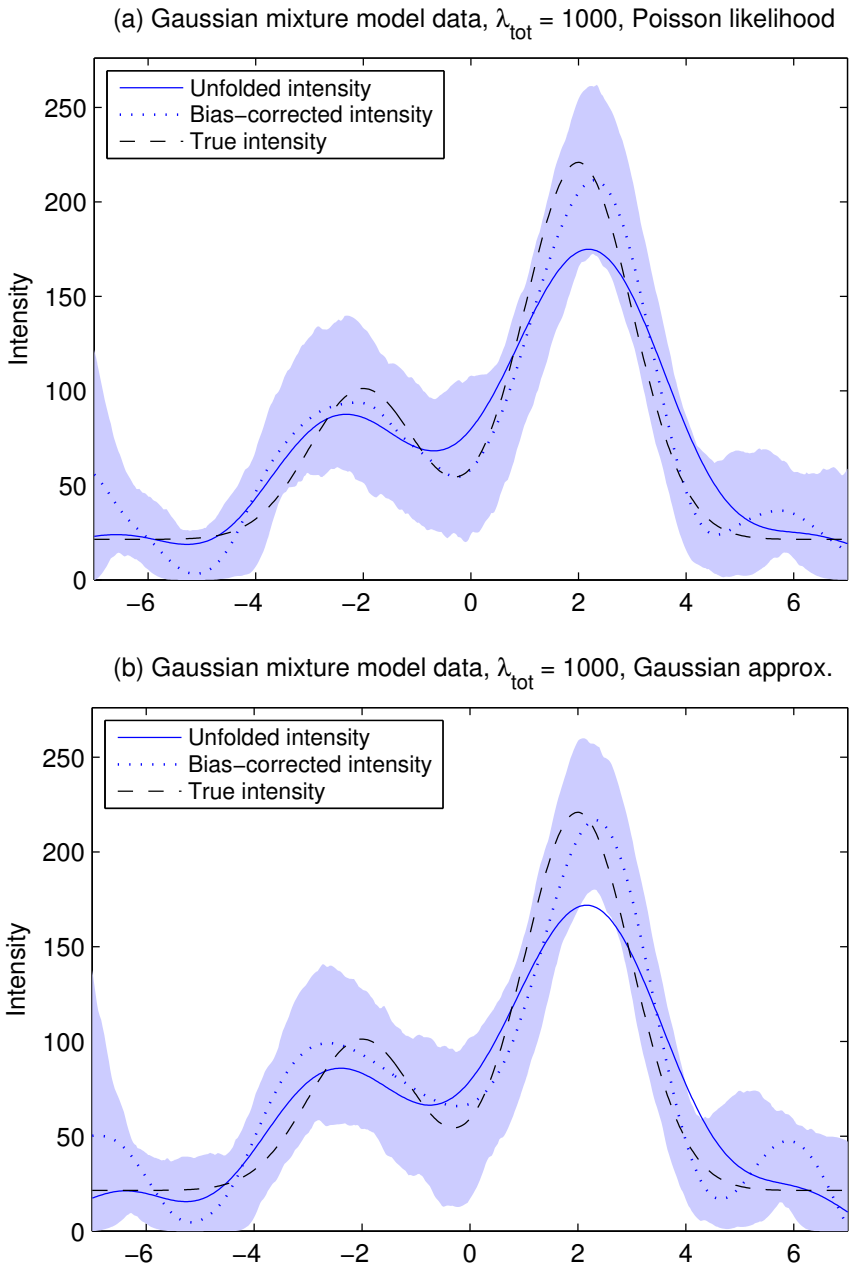

FIG 4. *Comparison of unfolding results obtained using (a) the full Poisson likelihood and (b) a Gaussian approximation to the full likelihood. The sample size was* $\lambda_{\text{tot}} = 1\,000$ *and the confidence intervals are the 95 % iteratively bias-corrected intervals obtained using* $N_{\text{BC}} = 15$ *bias correction iterations.*

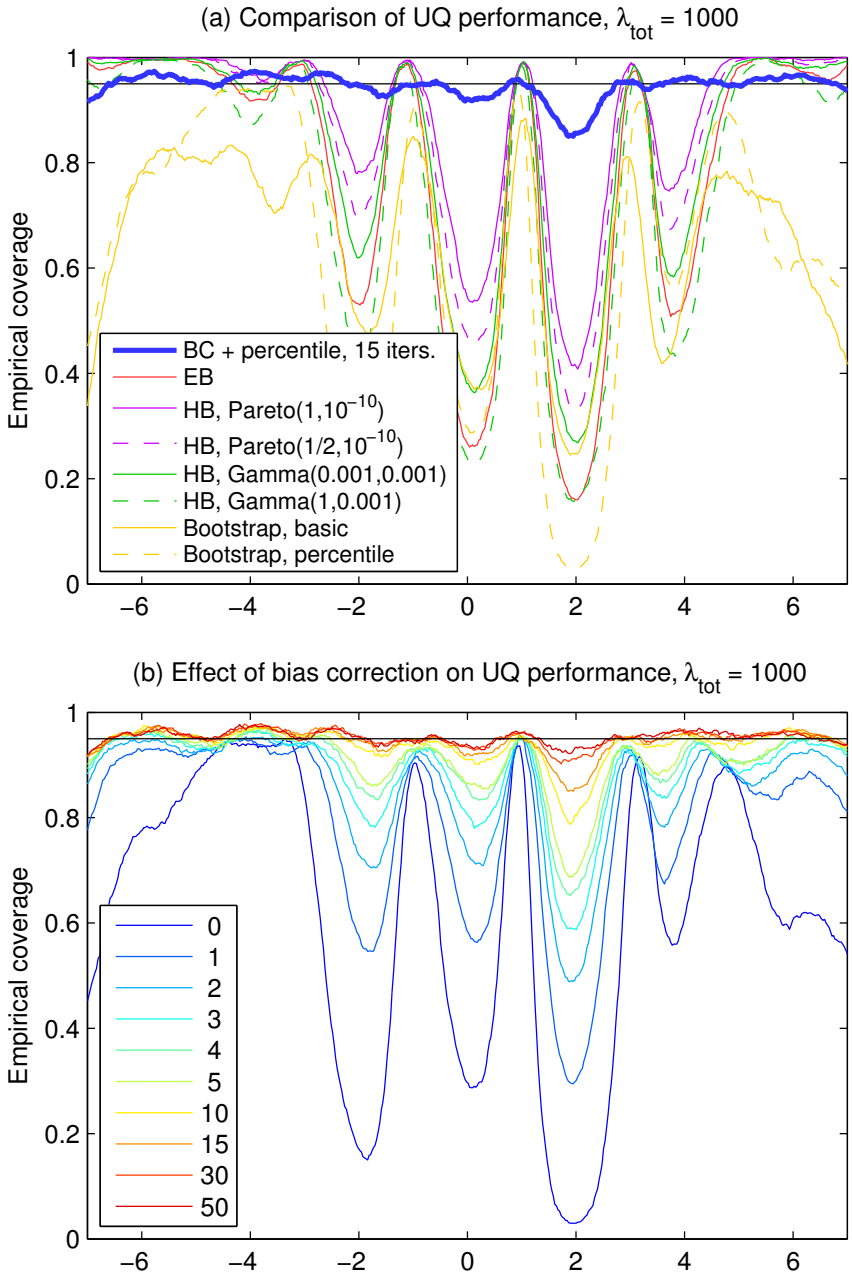

FIG 5. *Coverage studies with* $\lambda_{\text{tot}} = 1\,000$. *Figure (a) compares the empirical coverage of the iteratively bias-corrected intervals with 15 bias correction iterations to that of empirical Bayes (EB) and hierarchical Bayes (HB) credible intervals as well as the non-bias-corrected bootstrap percentile and basic intervals. Figure (b) shows the empirical coverage of the bias-corrected intervals as the number of bias-correction iterations is varied between 0 and 50. All intervals are formed for 95 % nominal coverage shown by the horizontal line.*

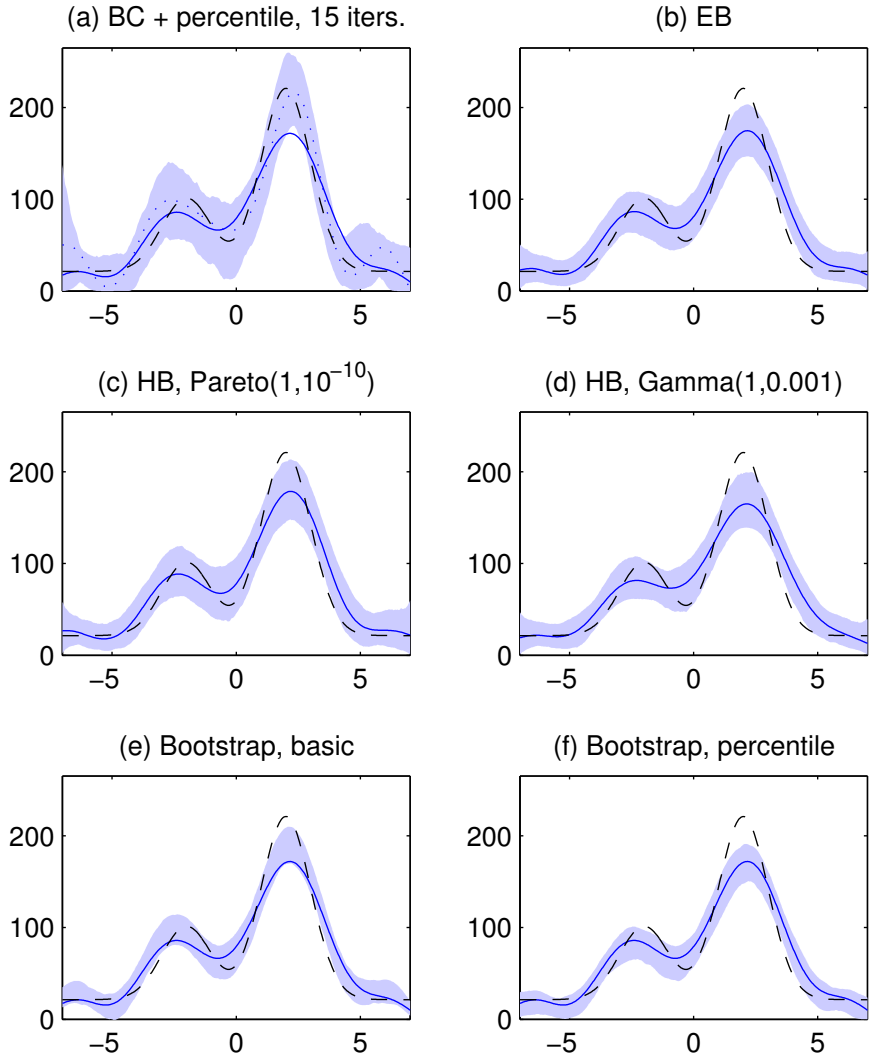

FIG 6. *One realization of the various 95 % confidence intervals with* $\lambda_{\text{tot}} = 1000$. *The intervals shown are (a) the iteratively bias-corrected intervals with 15 bias correction iterations, (b)–(d) credible intervals of the empirical Bayes (EB) and the two extremal hierarchical Bayes (HB) posteriors, (e) bootstrap basic intervals and (f) bootstrap percentile intervals. Figures (a), (e) and (f) were computed using the Gaussian approximation to the Poisson likelihood. Also shown are the corresponding unfolded point estimates* $\hat{f}$ *(solid lines) and the true intensity* f *(dashed lines). In the case of Figure (a), also the bias-corrected point estimate* $\hat{f}_{\text{BC}}$ *is given (dotted line).*

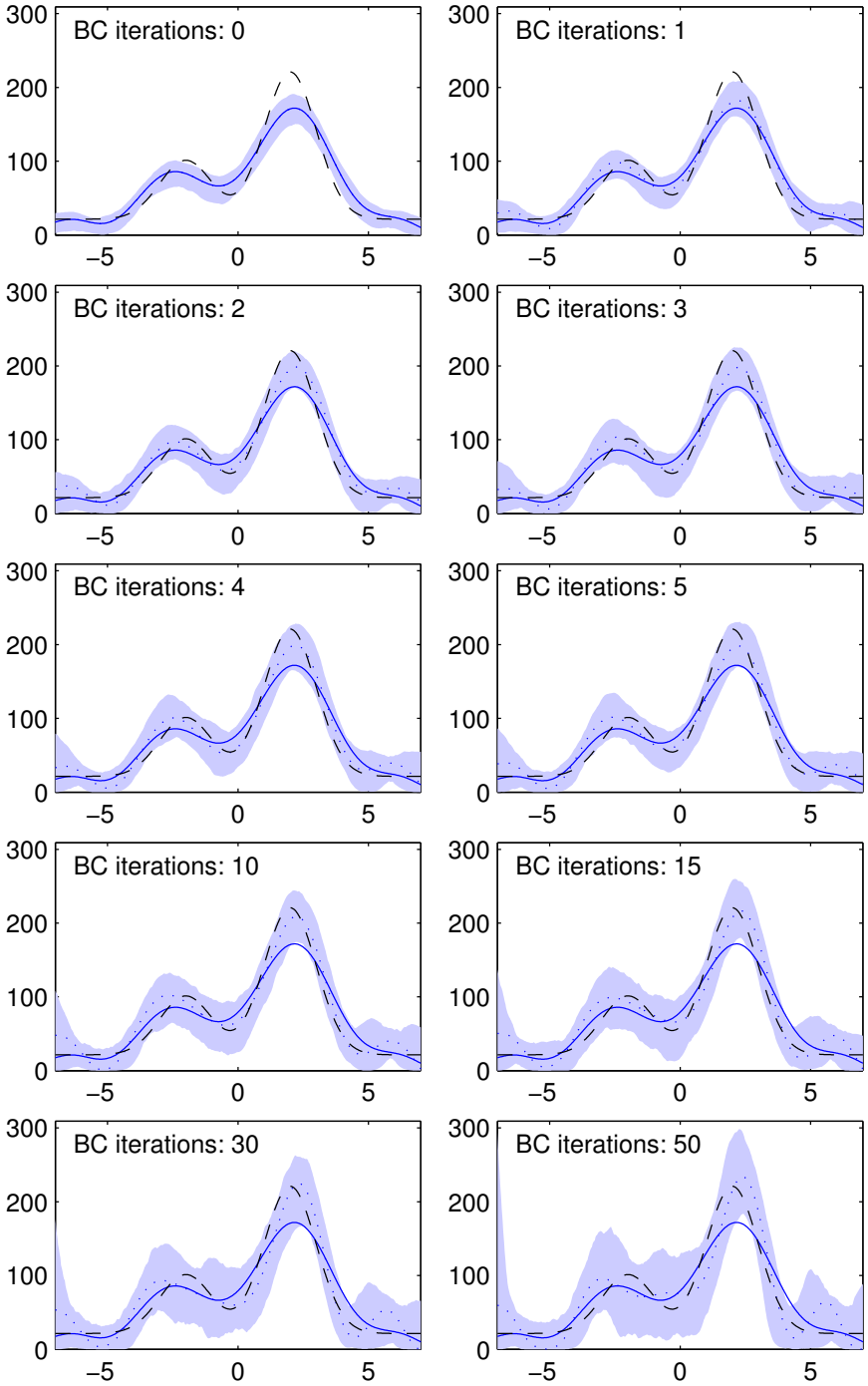

FIG 7. *Comparison of the 95 % iteratively bias-corrected intervals with varying amounts of bias correction with* $\lambda_{\text{tot}} = 1\,000$. *Also shown are the true intensity* f *(dashed lines), the unfolded point estimate* $\hat{f}$ *(solid lines) and the various bias-corrected point estimates* $\hat{f}_{\text{BC}}$ *(dotted lines). All the estimates were computed using the Gaussian approximation to the Poisson likelihood.*

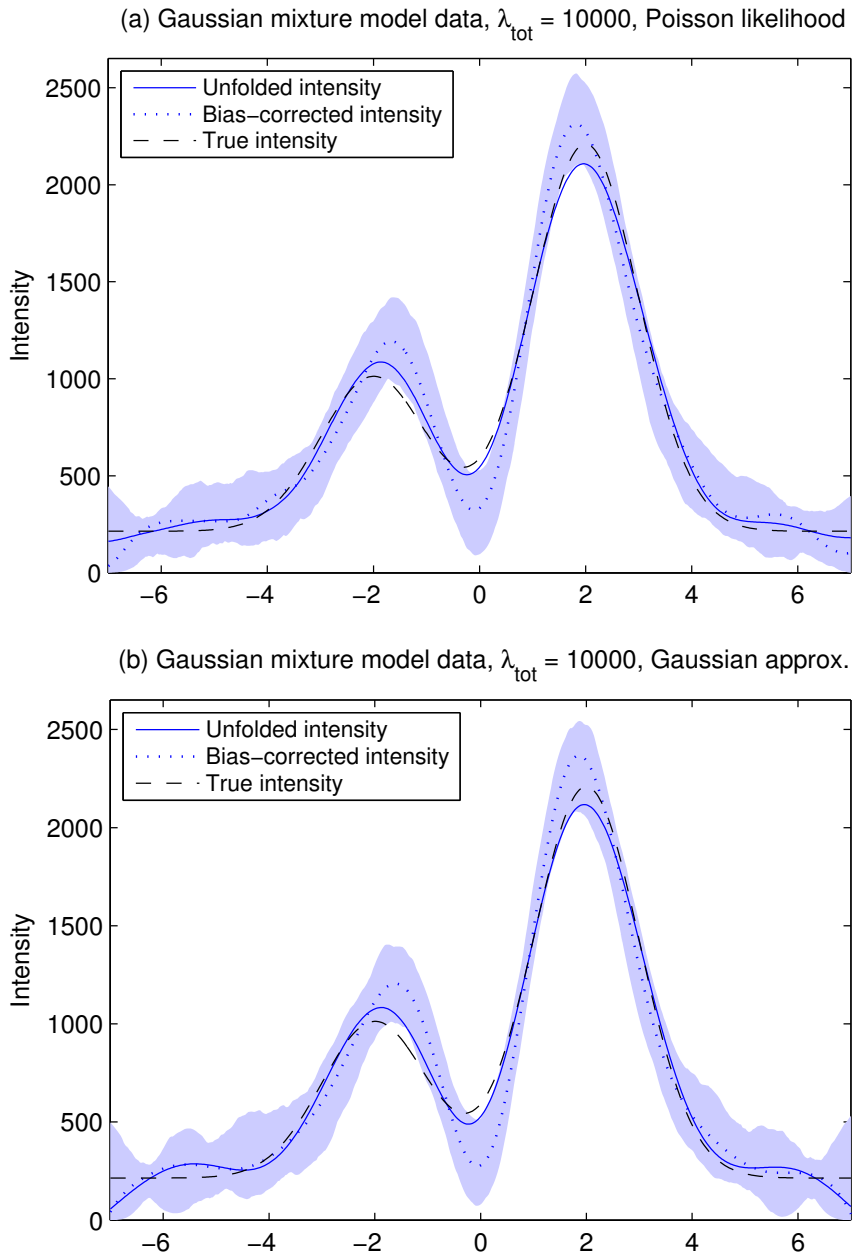

FIG 8. *Comparison of unfolding results obtained using (a) the full Poisson likelihood and (b) a Gaussian approximation to the full likelihood. The sample size was* $\lambda_{\text{tot}} = 10\,000$ *and the confidence intervals are the 95 % iteratively bias-corrected intervals obtained using* $N_{\text{BC}} = 5$ *bias correction iterations.*

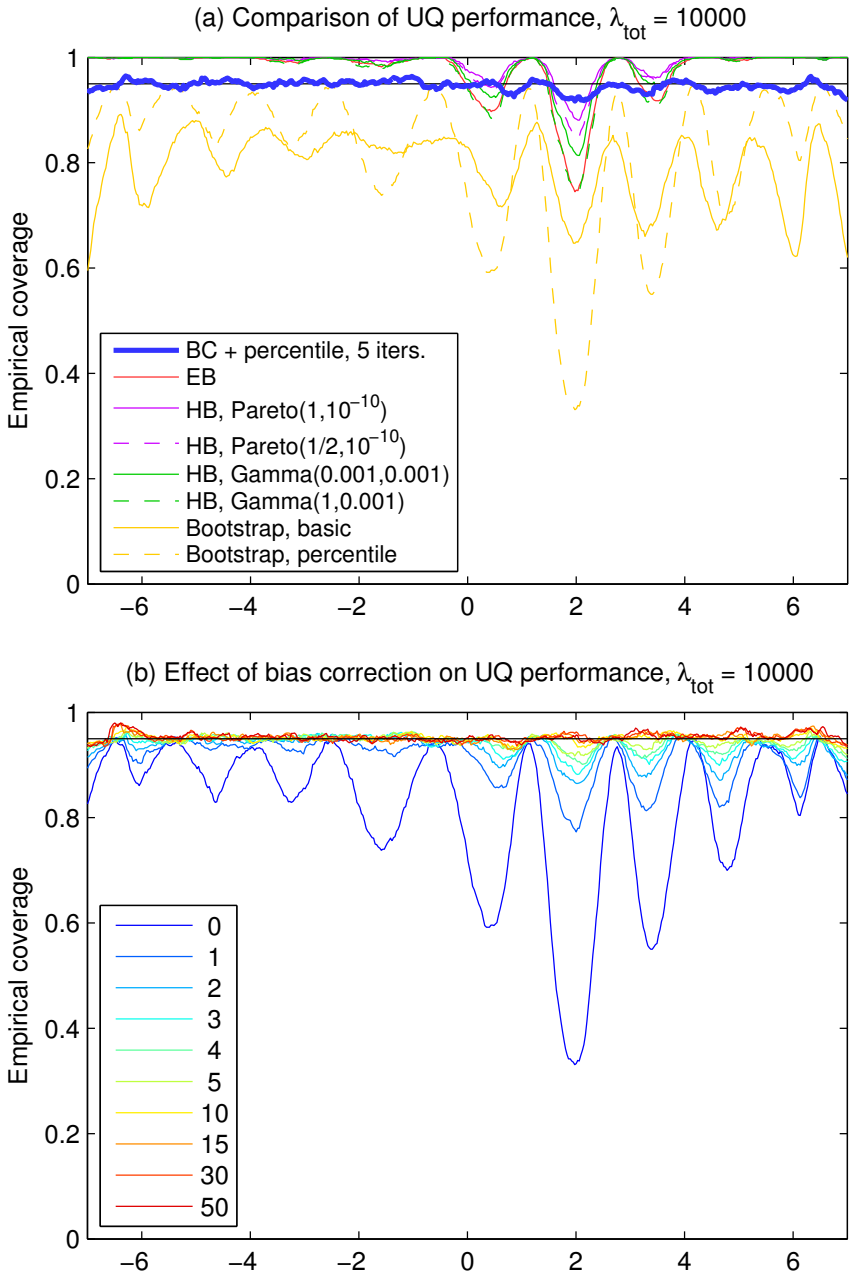

FIG 9. *Coverage studies with* $\lambda_{\text{tot}} = 10\,000$. *Figure (a) compares the empirical coverage of the iteratively bias-corrected intervals with 5 bias correction iterations to that of empirical Bayes (EB) and hierarchical Bayes (HB) credible intervals as well as the non-bias-corrected bootstrap percentile and basic intervals. Figure (b) shows the empirical coverage of the bias-corrected intervals as the number of bias-correction iterations is varied between 0 and 50. All intervals are formed for 95 % nominal coverage shown by the horizontal line.*

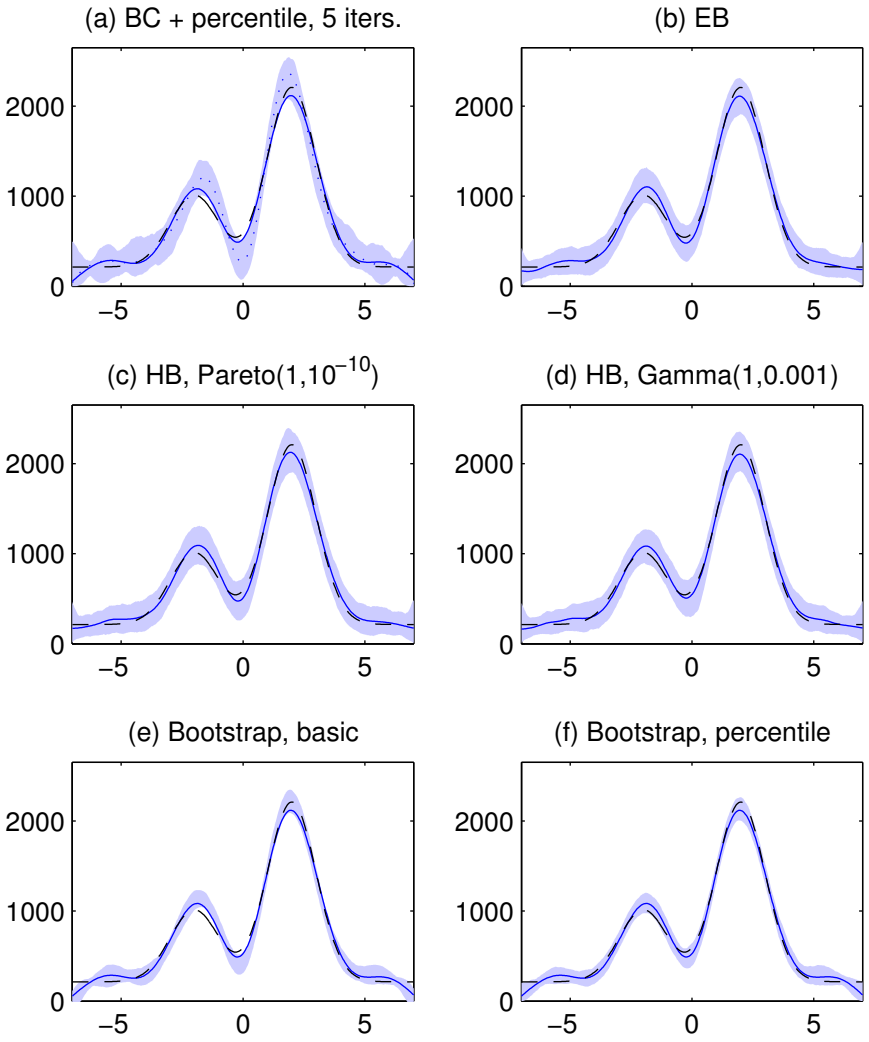

FIG 10. *One realization of the various 95 % confidence intervals with* $\lambda_{\text{tot}} = 10\,000$. *The intervals shown are (a) the iteratively bias-corrected intervals with 5 bias correction iterations, (b)–(d) credible intervals of the empirical Bayes (EB) and the two extremal hierarchical Bayes (HB) posteriors, (e) bootstrap basic intervals and (f) bootstrap percentile intervals. Figures (a), (e) and (f) were computed using the Gaussian approximation to the Poisson likelihood. Also shown are the corresponding unfolded point estimates* $\hat{f}$ *(solid lines) and the true intensity* f *(dashed lines). In the case of Figure (a), also the bias-corrected point estimate* $\hat{f}_{\text{BC}}$ *is given (dotted line).*

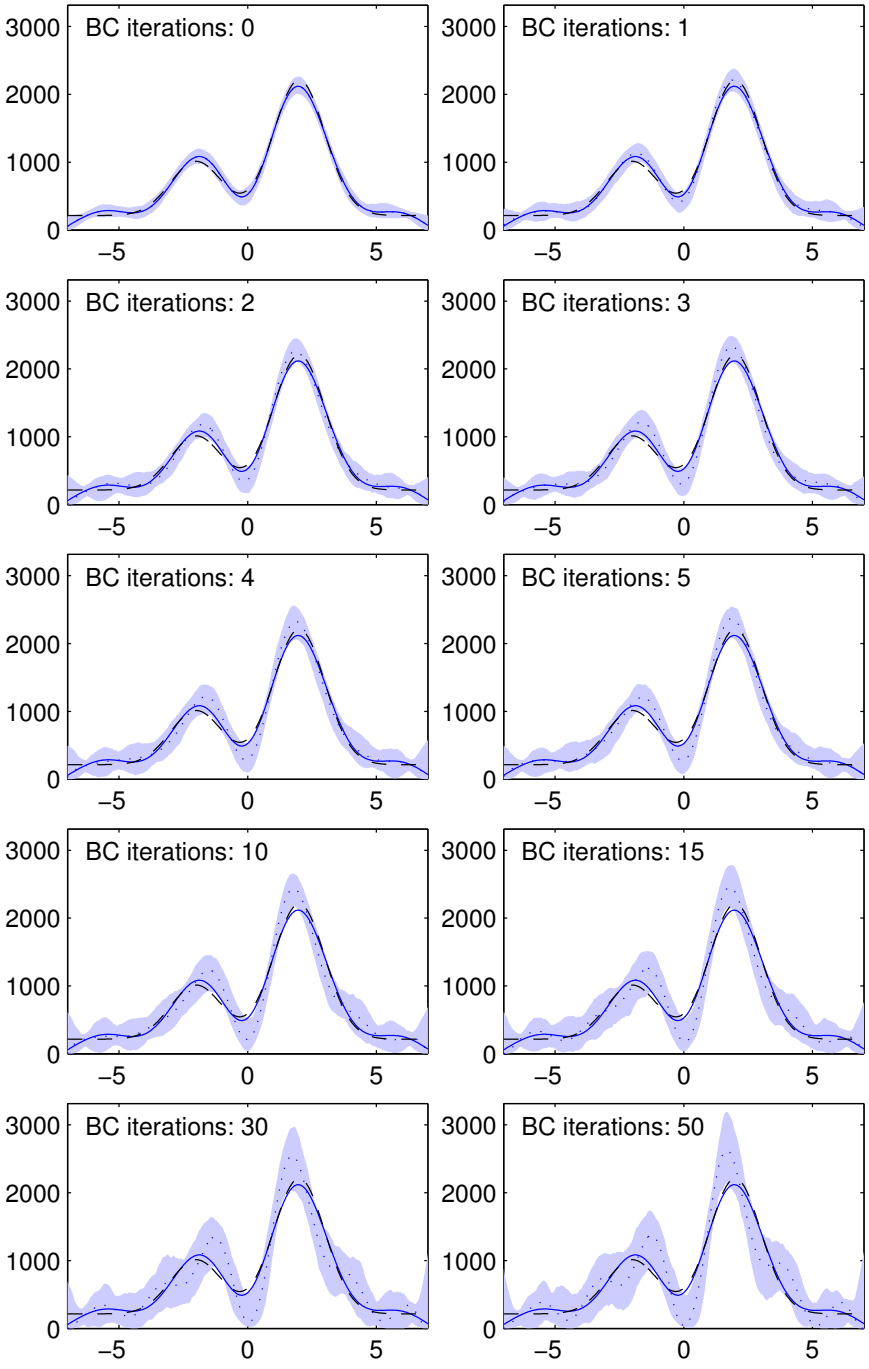

FIG 11. *Comparison of the 95 % iteratively bias-corrected intervals with varying amounts of bias correction with* $\lambda_{\text{tot}} = 10000$. *Also shown are the true intensity* f *(dashed lines), the unfolded point estimate* $\hat{f}$ *(solid lines) and the various bias-corrected point estimates* $\hat{f}_{\text{BC}}$ *(dotted lines). All the estimates were computed using the Gaussian approximation to the Poisson likelihood.*

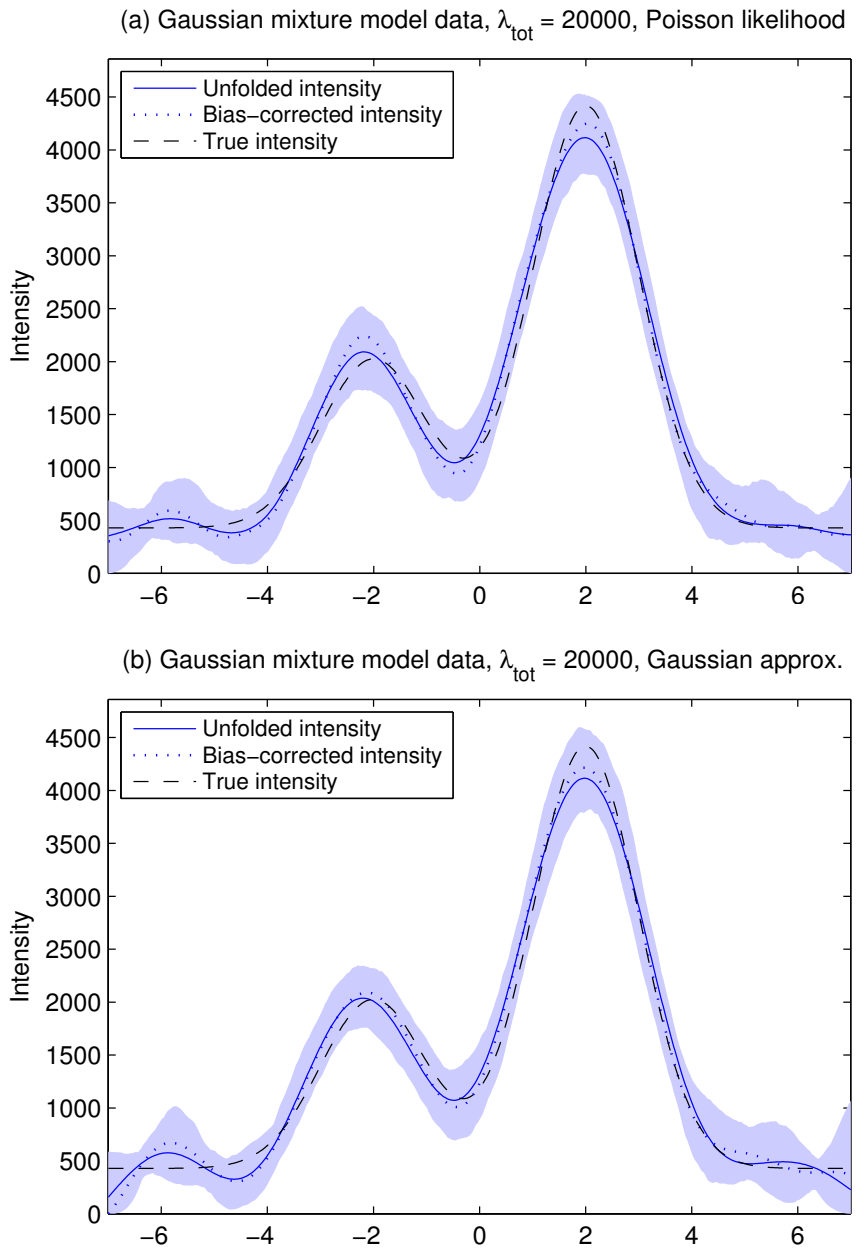

FIG 12. *Comparison of unfolding results obtained using (a) the full Poisson likelihood and (b) a Gaussian approximation to the full likelihood. The sample size was* $\lambda_{\text{tot}} = 20\,000$ *and the confidence intervals are the 95 % iteratively bias-corrected intervals obtained using* $N_{\text{BC}} = 5$ *bias correction iterations.*

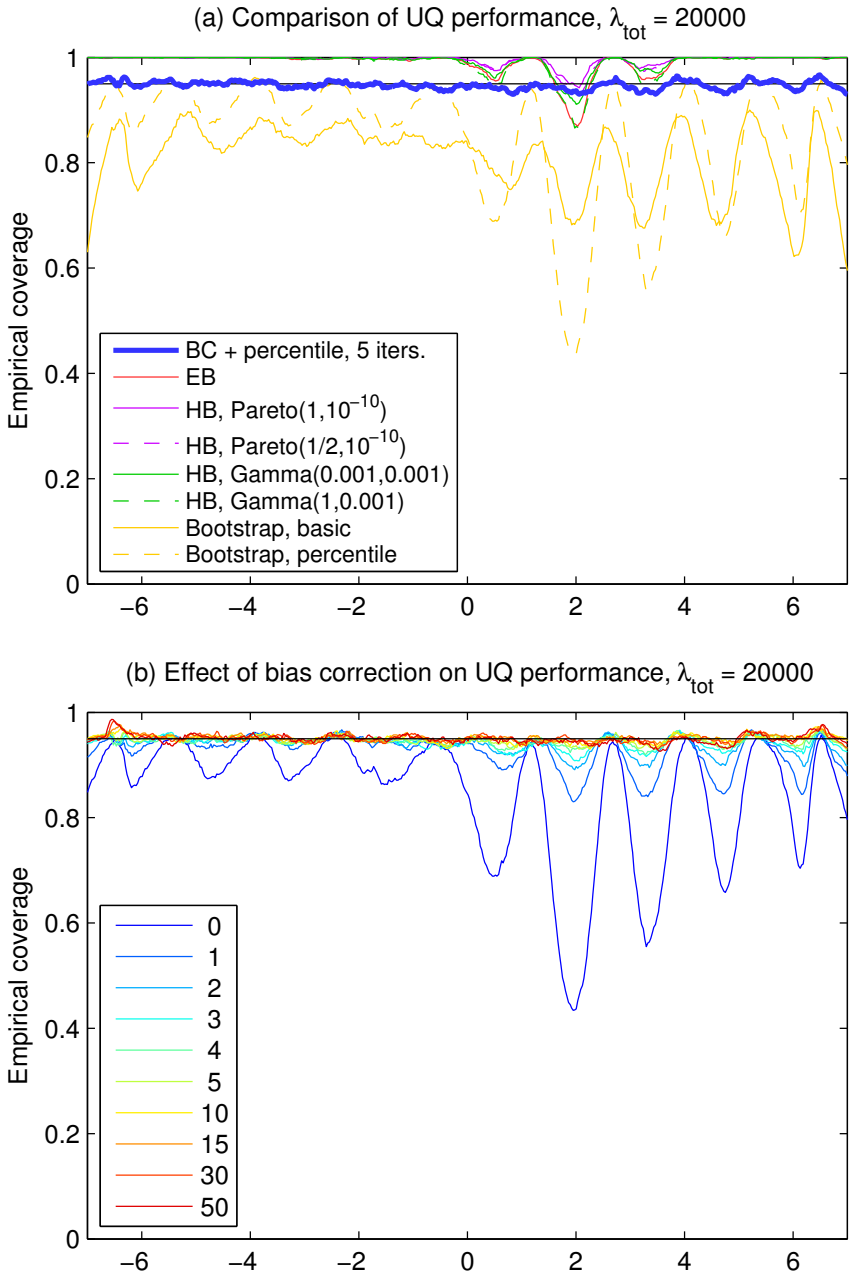

FIG 13. *Coverage studies with* $\lambda_{\text{tot}} = 20000$. *Figure (a) compares the empirical coverage of the iteratively bias-corrected intervals with 5 bias correction iterations to that of empirical Bayes (EB) and hierarchical Bayes (HB) credible intervals as well as the non-bias-corrected bootstrap percentile and basic intervals. Figure (b) shows the empirical coverage of the bias-corrected intervals as the number of bias-correction iterations is varied between 0 and 50. All intervals are formed for 95 % nominal coverage shown by the horizontal line.*

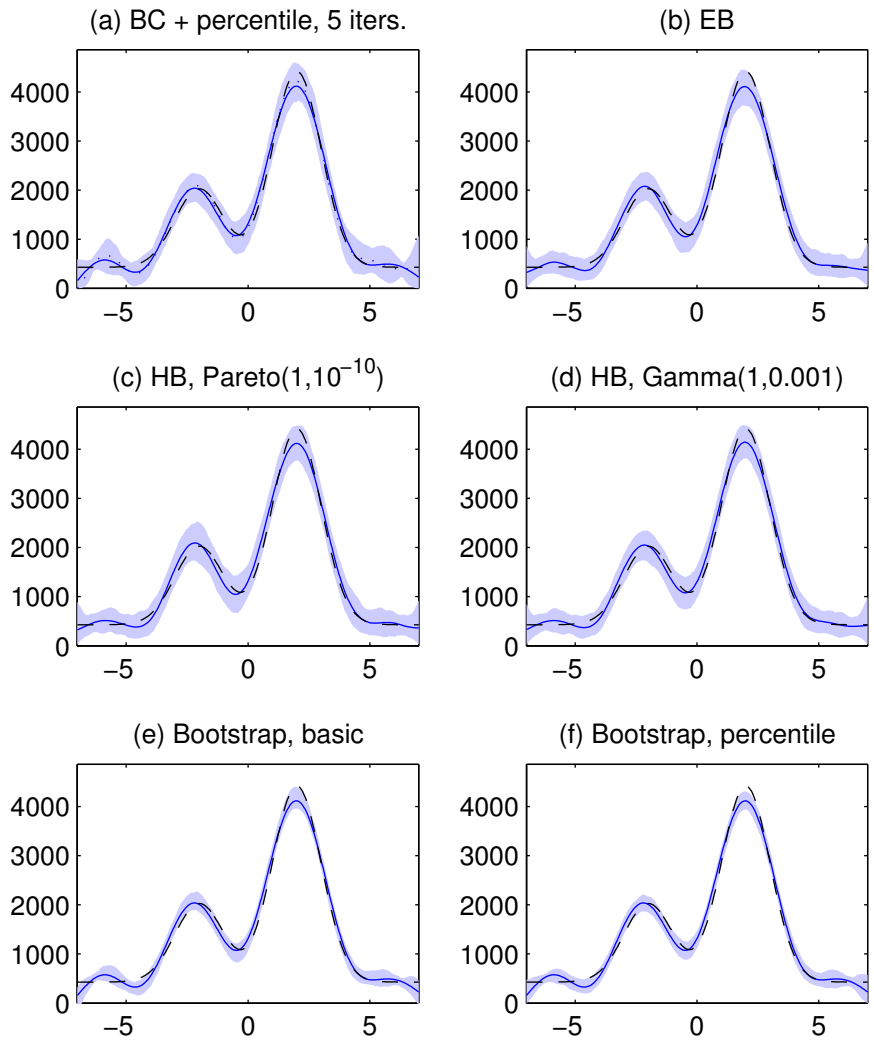

FIG 14. *One realization of the various 95 % confidence intervals with* $\lambda_{\text{tot}} = 20\,000$. *The intervals shown are (a) the iteratively bias-corrected intervals with 5 bias correction iterations, (b)–(d) credible intervals of the empirical Bayes (EB) and the two extremal hierarchical Bayes (HB) posteriors, (e) bootstrap basic intervals and (f) bootstrap percentile intervals. Figures (a), (e) and (f) were computed using the Gaussian approximation to the Poisson likelihood. Also shown are the corresponding unfolded point estimates* $\hat{f}$ *(solid lines) and the true intensity* f *(dashed lines). In the case of Figure (a), also the bias-corrected point estimate* $\hat{f}_{\text{BC}}$ *is given (dotted line).*

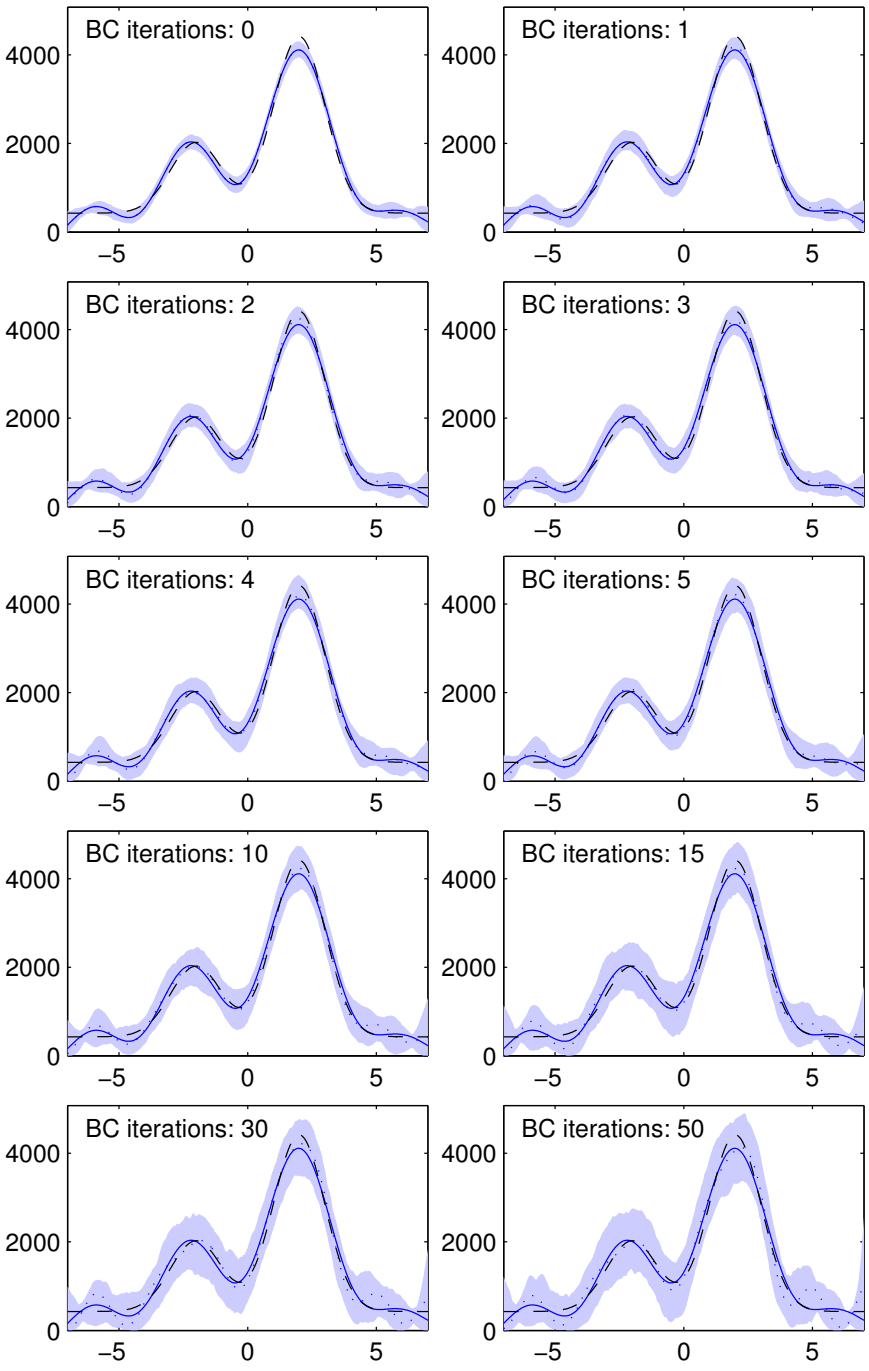

FIG 15. *Comparison of the 95 % iteratively bias-corrected intervals with varying amounts of bias correction with* $\lambda_{\text{tot}} = 20000$. *Also shown are the true intensity* f *(dashed lines), the unfolded point estimate* $\hat{f}$ *(solid lines) and the various bias-corrected point estimates* $\hat{f}_{\text{BC}}$ *(dotted lines). All the estimates were computed using the Gaussian approximation to the Poisson likelihood.*

References.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1** 473–514.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **1** 515–534.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 473–511.

Geyer, C. J. and Johnson, L. T. (2013). `mcmc`: Markov chain Monte Carlo. R package, version 0.9-2, available at CRAN.

Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician* **52** 93–100.

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press.

Section de Mathématiques
École Polytechnique Fédérale de Lausanne
EPFL Station 8, 1015 Lausanne
Switzerland
E-mail: mikael.kuusela@epfl.ch
victor.panaretos@epfl.ch