SIAM J. SCI. COMPUT.
Vol. 42, No. 2, pp. A1269–A1288

MCMC ALGORITHMS FOR COMPUTATIONAL UQ OF NONNEGATIVITY CONSTRAINED LINEAR INVERSE PROBLEMS[*]

JOHNATHAN M. BARDSLEY[†] AND PER CHRISTIAN HANSEN[‡]

Abstract. In many inverse problems, a nonnegativity constraint is natural. Moreover, in some cases, we expect the vector of unknown parameters to have zero components. When a Bayesian approach is taken, this motivates a desire for prior probability density (and hence posterior probability density) functions that have positive mass at the boundary of the set $\{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{x} \geq \mathbf{0}\}$. Unfortunately, it is difficult to define a prior with this property that yields computationally tractable inference for large-scale inverse problems. In this paper, we use nonnegativity constrained optimization to define such prior and posterior density functions when the measurement error is either Gaussian or Poisson distributed. The numerical optimization methods we use are highly efficient, and hence our approach is computationally tractable even in large-scale cases. We embed our nonnegativity constrained optimization approach within a hierarchical framework, obtaining Gibbs samplers for both Gaussian and Poisson distributed measurement cases. Finally, we test the resulting Markov chain Monte Carlo methods on examples from both image deblurring and positron emission tomography.

Key words. inverse problems, uncertainty quantification, Bayesian methods, Markov chain Monte Carlo, nonnegativity constraints

AMS subject classifications. 15A29, 62F15, 65C05, 65C40, 65F22, 94A08

DOI. 10.1137/18M1234588

1. Introduction.

This work considers computational aspects of uncertainty quantification (UQ) for nonnegativity constrained linear inverse problems, formulated in a Bayesian setting which is natural for UQ. Our basis is the regression problem with nonnegativity constraints,

$$\min_{\boldsymbol{x}} G(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \geq \mathbf{0}\ , \tag{1.1}$$

where $G(\boldsymbol{x})$ is a goodness-of-fit measure suited for the specific data—e.g., a (weighted) 2-norm of the residual for data with Gaussian noise. Nonnegativity constraints appear in many applications where the underlying physics dictates that the solution cannot be negative. This is the case, e.g., for absorption coefficients in computed tomography [8], image intensities in astronomical imaging [28], and wave velocities in seismic travel-time tomography [18, Chapter 17]. Hence, nonnegativity constraints are very important for producing meaningful reconstructions, and there is a need for computational UQ methods that handle these constraints efficiently and in a rigorous manner.

In the Bayesian framework, all information, such as the priors on the solution, are expressed in terms of probability density functions. If we replace the nonnegativity

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section December 20, 2018; accepted for publication (in revised form) January 16, 2020; published electronically April 27, 2020.
https://doi.org/10.1137/18M1234588
Funding: The work of the first author was partially supported by the Danish Research Council for Independent Research—Natural Science, grant 4002-00123. The work of the second author was supported by a Villum Investigator grant 25893 from the Villum Foundation.

[†] Department of Mathematical Sciences, University of Montana, Missoula, MT 59812 (bardsleyj@mso.umt.edu).

[‡] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kangens, Lyngby DK-2800, Denmark (pcha@dtu.dk).

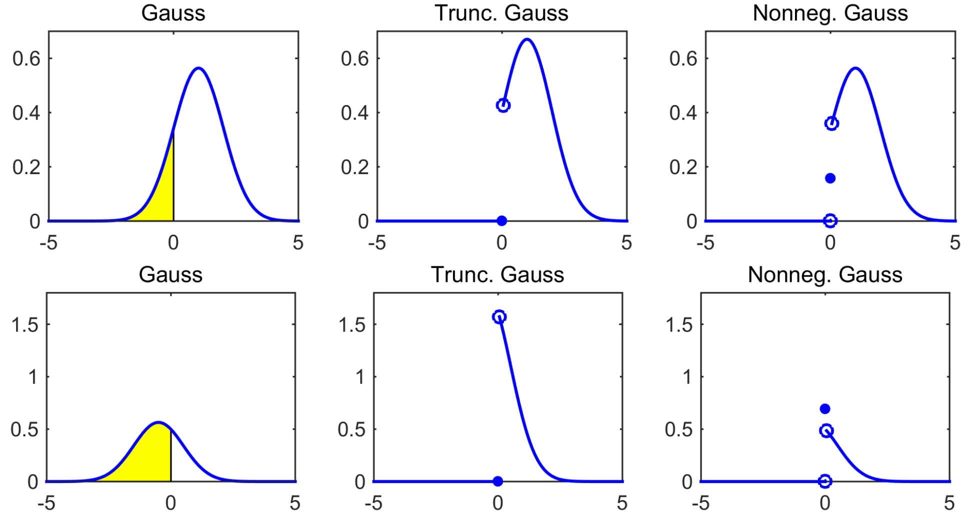

FIG. 1. *Incorporating a nonnegativity constraint into a Gaussian distribution* $p(x)$. *Left: (top) a Gaussian with mean* 1 *and variance* 1 *and (bottom) a Gaussian with mean* $-1/2$ *and variance* 1 *the area below the curves for* $x \in [-\infty, 0]$ *(the shaded area) is denoted* I_0. *Middle: the truncated Gaussians obtained by setting* $p(x) = 0$ *for* $x \in [-\infty, 0]$ *and normalizing. Right: the "nonnegative Gaussians," which have probabilities* 0 *for* $x < 0$, I_0 *at* $x = 0$, *and* $p(x)$ *for* $x > 0$.

constraint in (1.1) with a positivity constraint, $\boldsymbol{x} > \mathbf{0}$, in the linear Gaussian case, a truncated Gaussian distribution results, and standard statistical techniques can be used for performing UQ; see, e.g., [4, 10]. In this work we specifically address situations in which we expect some of the components of $\boldsymbol{x}$ to equal zero—in fact, in some imaging problems a substantial fraction of the solutions elements/pixels may be zero.

Figure 1 illustrates the "nonnegative Gaussian" probability density function, $p_{\mathrm{NN}}(x)$, that we use in this work. Also plotted in the figure is the truncated Gaussian probability density function mentioned in the previous paragraph. For insight, we construct these two distributions from the same underlying Gaussian distribution, $p(x)$, and compare them. Note that they both have probability zero for $x < 0$ and are proportional to $p(x)$ for $x > 0$, but the truncated Gaussian has probability zero at $x = 0$, whereas $p_{\mathrm{NN}}(0) = \int_{-\infty}^{0} p(x)\,dx$. The nonnegative Gaussian is, therefore, attractive in applications where $x = 0$ with positive probability.

In this paper, we make use of a nonnegative multivariate Gaussian distribution defined through a nonnegativity constrained least squares problem. Unfortunately, except for very simple forms of the Gaussian random vector, the corresponding probability density function does not have a convenient analytical expression suited for numerical computations. Our approach follows that in [4], where a Markov chain Monte Carlo (MCMC) method was used to sample the constrained solution's probability distribution in the Gaussian measurement error case. However, in [4], there was a lack of rigorous mathematical justification for the approach. Thus, we present a rigorous justification of the algorithm from [4] in this paper. We then extend the approach to the Poisson measurement error case by embedding a nonnegativity constrained convex optimization problem within an analogous Gibbs sampler.

The key idea in our algorithm is to solve a nonnegativity constrained stochastic least squares problem to obtain constrained samples as one step within a Gibbs sampler. This approach works well and provides different results than are obtained

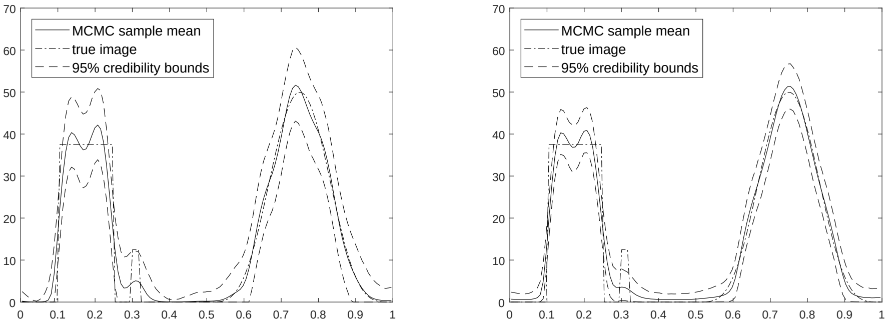

Fig. 2. *Results obtained for a one-dimensional image deblurring test case, using the nonnegative Gaussian (left) and the truncated Gaussian (right).*

using a truncated Gaussian, as can be seen in Figure 2, where the two approaches are implemented on the one-dimensional image deblurring test case presented in section 4. In this example, the nonnegative Gaussian (plotted on the left) provides a better reconstruction in the lower-intensity regions of the image; in particular, note the reconstruction of the small spike. This is likely due to the fact that the nonnegative Gaussian is a more accurate model near where the true signal is zero.

MCMC sampling with inequality constraints has been studied by other researchers. The work of Michalak and collaborators [21, 22], for example, makes use of truncated Gaussian priors. Their approach has the benefit that conditional densities are formulated for each parameter, which are in turn used to define the posterior density function and a natural componentwise Gibbs sampler. The downside of this and similar approaches is that componentwise Gibbs samplers are slow to converge for large-scale problems. Calvetti and collaborators (see [9]) gave an algorithm with improved efficiency by using the hit-and-run, or random direction, variant of the Gibbs sampler with direction biased according to the Cholesky factor of the covariance of the unconstrained Gaussian distribution. However, this improvement is computationally intensive for high-dimensional problems, and it samples from a truncated Gaussian with zero mass at the constraint boundary. A third option is to assume a prior that has support on $\boldsymbol{x} > \mathbf{0}$. For example, one could assume a Gamma or inverse-Gamma prior for the components of $\boldsymbol{x}$ [12], which can be defined so that sparsity is enforced in $\boldsymbol{x}$. Another option, which has also been used in the context of inverse problems (see, e.g., [2, 29]), is to assume a lognormal prior for $\boldsymbol{x}$; then $\mathbf{z} = \ln(\boldsymbol{x})$ is normally distributed on $\mathbb{R}^N$ and $\boldsymbol{x} = \exp(\mathbf{z}) > \mathbf{0}$.

Our paper is organized as follows. In section 2, we focus on the Gaussian measurement case, presenting a full hierarchical model and the nonnegativity constrained hierarchical Gibbs sampler. In section 3, we proceed in the same fashion but begin with the assumption of Poisson distributed measurements. This results in changes to the hierarchical model and the nonnegative hierarchical Gibbs sampler. We conclude with a variety of numerical examples in section 4.

2. Gaussian measurement error.

In this section, we consider additive Gaussian statistical models of the form

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\boldsymbol{I}_M), \tag{2.1}$$

where $\boldsymbol{b} \in \mathbb{R}^M$ is the vector of measurements, $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ is the forward model matrix, $\boldsymbol{x} \in \mathbb{R}^N$ is the vector of unknown parameters, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\boldsymbol{I}_M)$ means that $\boldsymbol{\epsilon}$ is an M-dimensional Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\lambda^{-1}\boldsymbol{I}_M$, with $\boldsymbol{I}_M$ denoting the $M \times M$ identity matrix.

2.1. A Bayesian hierarchical model and Gibbs sampler. The random vector $\boldsymbol{b}$ in (2.1) has probability density function

$$p(\boldsymbol{b}\,|\,\boldsymbol{x},\lambda) \propto \lambda^{M/2}\exp\left(-\frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2\right), \tag{2.2}$$

where $\propto$ denotes proportionality. The term $\lambda^{M/2}$ in (2.2) appears in the normalization constant of the Gaussian and is included because it will be important when we later define a hierarchical model. The likelihood function is given by (2.2) as a function of $\boldsymbol{x}$, and, as is well known, computing its maximizer is an ill-posed problem.

There are various methods to make the solution of inverse problems well-posed [17]. In this paper, we take the Bayesian approach [2], which requires the definition of a prior probability density function on $\boldsymbol{x}$. We make the assumption that the prior is Gaussian of the form $\boldsymbol{x}|\delta \sim \mathcal{N}\left(\mathbf{0}, (\delta\boldsymbol{L})^{-1}\right)$, which has probability density function

$$p(\boldsymbol{x}\,|\,\delta) \propto \delta^{\bar{N}/2}\exp\left(-\frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}\right), \tag{2.3}$$

where $\bar{N} = \mathrm{rank}(\boldsymbol{L})$ and if $\bar{N} < N$, $p(\boldsymbol{x}|\delta)$ is known as an intrinsic Gaussian [27]. As in (2.2), we include the term $\delta^{\bar{N}/2}$ from the normalization constant of the Gaussian because it will be important when we define the hierarchical model. We limit ourselves to Gaussian priors because, as we will see next, they result in Gaussian posteriors, which are easy to work with even in high-dimensional cases. However, Gaussian priors are also flexible and can incorporate problem-specific prior information, especially if Gaussian Markov random fields are used; see, e.g., [2, 27].

Once a prior and a likelihood have been chosen, using Bayes' law we multiply them together to obtain the posterior density function:

$$\begin{aligned} p(\boldsymbol{x}\,|\,\boldsymbol{b},\lambda,\delta) &\propto p(\boldsymbol{b}\,|\,\boldsymbol{x},\lambda)\,p(\boldsymbol{x}\,|\,\delta) \\ &\propto \lambda^{M/2}\,\delta^{\bar{N}/2}\exp\left(-\frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2-\frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}\right). \end{aligned} \tag{2.4}$$

Independent samples from the posterior (2.4) can be computed using the generative model

$$\begin{gathered} \boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}} = \arg\min_{\boldsymbol{x}} \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{x}-\hat{\boldsymbol{b}}\|^2 + \frac{\delta}{2}\|\boldsymbol{L}^{1/2}\boldsymbol{x}-\hat{\mathbf{c}}\|^2, \\ \text{where } \hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b},\lambda^{-1}\boldsymbol{I}_M) \text{ and } \hat{\mathbf{c}} \sim \mathcal{N}(\mathbf{0},\delta^{-1}\boldsymbol{I}_N), \end{gathered} \tag{2.5}$$

where $\boldsymbol{L} = \boldsymbol{L}^{T/2}\boldsymbol{L}^{1/2}$ is a Cholesky factorization or some other matrix square root. To see that (2.5) yields samples from (2.4), note that the normal equations of (2.5) have solution

$$\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}} = (\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}\left(\lambda\boldsymbol{A}^T\hat{\boldsymbol{b}}+\delta\boldsymbol{L}^{T/2}\hat{\mathbf{c}}\right),$$

which is a Gaussian distribution with mean $(\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}\lambda\boldsymbol{A}^T\boldsymbol{b}$ and covariance matrix $(\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}$, just as is the case for (2.4). To guarantee that $\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L}$ is invertible, we assume $N(\boldsymbol{A})\cap N(\boldsymbol{L}) = \{\mathbf{0}\}$, where $N(\cdot)$ denotes null space. There are alternative ways of expressing (2.5), but we use this formulation because it serves as motivation for the stochastic optimization problems used in *both* the Gaussian and the Poisson measurement error cases.

In [3], another level is added to the Bayesian statistical model by assuming Gamma distributions on the hyperparameters λ and δ:

$$p(\lambda) \propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda), \quad \lambda > 0, \tag{2.6}$$

$$p(\delta) \propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta), \quad \delta > 0. \tag{2.7}$$

Gamma distributions are chosen because they yield conditional distributions that are also Gamma distributed—a property known as conjugacy—making it straightforward to define the Gibbs sampler below. The choice of the hyperparameters $(\alpha_\lambda, \beta_\lambda)$ and $(\alpha_\delta, \beta_\delta)$ is important; specifically, they should be chosen so that the hyperpriors have minimal influence, with respect to λ and δ, on the posterior. We have used $\alpha_\lambda = \alpha_\beta = 1$ and $\beta_\lambda = \beta_\delta = 10^{-4}$ on a wide range of examples and have found these choices to be robust.

Taken together, (2.2), (2.3), (2.6), and (2.7) constitute a *Bayesian hierarchical model*. The resulting posterior density function, obtained once again using Bayes' law, is given by

$$\begin{aligned} p(\boldsymbol{x}, \lambda, \delta \,|\, \boldsymbol{b}) &\propto p(\boldsymbol{b} \,|\, \boldsymbol{x}, \lambda) p\left(\boldsymbol{x} \,|\, \delta\right) p(\lambda)\, p(\delta) \\ &= \lambda^{M/2 + \alpha_\lambda - 1}\, \delta^{\bar{N}/2 + \alpha_\delta - 1} \\ &\qquad \times \exp\left(-\frac{\lambda}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \frac{\delta}{2} \boldsymbol{x}^T \boldsymbol{L} \boldsymbol{x} - \beta_\lambda \lambda - \beta_\delta \delta \right), \quad \lambda, \delta > 0. \end{aligned} \tag{2.8}$$

The hierarchical Gibbs sampler of [3, 2] is obtained by cyclically sampling from the conditional distributions $p(\lambda, \delta \,|\, \boldsymbol{b}, \boldsymbol{x})$ and $p(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda, \delta)$.

Hierarchical Gibbs Sampler

```
Initialization: Choose $(\lambda_0, \delta_0, \boldsymbol{x}^0)$.
For $k = 1, 2, \ldots$
    Compute $(\lambda_k, \delta_k) \sim p(\lambda, \delta | \boldsymbol{b}, \boldsymbol{x}^{k-1})$ as follows:
        Compute $\lambda_k \sim \Gamma\left(M/2 + \alpha_\lambda, \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^{k-1} - \boldsymbol{b}\|^2 + \beta_\lambda\right)$.
        Compute $\delta_k \sim \Gamma\left(\overline{N}/2 + \alpha_\delta, \frac{1}{2}(\boldsymbol{x}^{k-1})^T \boldsymbol{L}\boldsymbol{x}^{k-1} + \beta_\delta\right)$.
    Compute $\boldsymbol{x}^k \sim p(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda_k, \delta_k)$ by solving (2.5) with $(\lambda, \delta) = (\lambda_k, \delta_k)$.
End
```

Remarks. The MCMC chain generated by hierarchical Gibbs converges in distribution to the full posterior $p(\boldsymbol{x}, \lambda, \delta \,|\, \boldsymbol{b})$ defined by (2.8); see [2, section 5.2] and the references therein. Thus, we say that $p(\boldsymbol{x}, \lambda, \delta \,|\, \boldsymbol{b})$ is the *target distribution* of hierarchical Gibbs. However, the question remains: how long does the MCMC chain need to be; i.e., how many iterations are required to ensure that the chain adequately represents the target distribution? In other words, if we perform K iterations, for what value of K has the MCMC chain $\{(\boldsymbol{x}^k, \lambda_k, \delta_k)\}_{k=1}^K$ converged in distribution to $p(\boldsymbol{x}, \lambda, \delta \,|\, \boldsymbol{b})$? Unfortunately, there are no definite tests for convergence in distribution (such as there are for numerical optimization methods), but there are a variety of diagnostics that provide evidence of convergence. In our numerical tests in section 4, we use the Geweke test for determining the convergence of our MCMC chains [13], but many alternatives exist, including the standard method shown in [12, section 11.4], as well as more recently developed methods, such as those found in [7, 14].

2.2. Nonnegativity constrained least squares sampling. In [4], a nonnegativity constraint is added to (2.5), yielding the nonnegativity constrained stochastic least squares problem

$$\begin{aligned}\boldsymbol{x}_{\lambda,\delta} &= \arg\min_{\boldsymbol{x}\geq\boldsymbol{0}} \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{x}-\hat{\boldsymbol{b}}\|^2 + \frac{\delta}{2}\|\boldsymbol{L}^{1/2}\boldsymbol{x}-\hat{\mathbf{c}}\|^2,\\ &\text{where } \hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b},\lambda^{-1}\boldsymbol{I}_M) \text{ and } \hat{\mathbf{c}} \sim \mathcal{N}(\boldsymbol{0},\delta^{-1}\boldsymbol{I}_N).\end{aligned} \tag{2.9}$$

As above, we assume $N(\boldsymbol{A}) \cap N(\boldsymbol{L}) = \{\boldsymbol{0}\}$ so that (2.9) is well-defined. We denote the probability density function for $\boldsymbol{x}_{\lambda,\delta}$ in (2.9) by $p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$. Recall that we are interested in imaging applications in which there is a positive probability that some components of $\boldsymbol{x}$ are zero, which is the case for (2.9). Moreover, highly efficient numerical methods exist for solving (2.9); we will use the active set method gradient projection conjugate gradient [23]. Thus, computing nonnegative samples using (2.9) is efficient, even for large-scale problems, and for this reason it is desirable to use.

In an attempt to derive a closed form for $p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$, we first consider the scalar case. Let $M = N = 1$; then (2.9) takes the form

$$\begin{aligned}x_{\lambda,\delta} &= \arg\min_{x\geq 0} \frac{\lambda}{2}(ax-\hat{b})^2 + \frac{\delta}{2}(\ell^{1/2}x-\hat{c})^2, \quad \hat{b}\sim\mathcal{N}(b,\lambda^{-1}) \text{ and } \hat{c}\sim\mathcal{N}(0,\delta^{-1})\\ &= \arg\min_{x\geq 0}\frac{1}{2}(\lambda a^2+\delta\ell)\left(x-x_{\lambda,\delta}^{\mathrm{UC}}\right)^2, \quad x_{\lambda,\delta}^{\mathrm{UC}} \sim \mathcal{N}\left(\frac{\lambda ab}{\lambda a^2+\delta\ell},(\lambda a^2+\delta\ell)^{-1}\right),\end{aligned} \tag{2.10}$$

where $x_{\lambda,\delta}^{\mathrm{UC}}$ is equivalently defined by (2.5). If we define $p_{\mathrm{UC}}(x|b,\lambda,\delta)$ and $p_{\mathrm{NN}}(x|b,\lambda,\delta)$ to be the probability density functions for $x_{\lambda,\delta}^{\mathrm{UC}}$ and $x_{\lambda,\delta}$, respectively, then it is straightforward to see from (2.10) that

$$p_{\mathrm{NN}}(x|b,\lambda,\delta) = \mathbf{1}_{\Omega}(x)\,p_{\mathrm{UC}}(x|b,\lambda,\delta) + \delta_0(x)\int_{-\infty}^{0} p_{\mathrm{UC}}(t|b,\lambda,\delta)dt, \tag{2.11}$$

where $\mathbf{1}_{\Omega}$ is the indicator function on $\Omega = \{x \in \mathbb{R} \,|\, x > 0\}$ and $\delta_0(x)$ is the delta function that equals 1 when $x = 0$ and 0 otherwise. Note that (2.11) is of the same type as the nonnegative Gaussian distributions plotted on the right in Figure 1.

In higher dimensions, a similar calculation yields the following equivalent formulation of (2.9):

$$\boldsymbol{x}_{\lambda,\delta} = \arg\min_{\boldsymbol{x}\geq\boldsymbol{0}} \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}}\right)^T(\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})\left(\boldsymbol{x}-\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}}\right), \tag{2.12}$$

where

$$\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}} \sim \mathcal{N}\left((\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}\lambda\boldsymbol{A}^T\boldsymbol{b},(\lambda\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}\right). \tag{2.13}$$

Note that (2.13) is equivalent to (2.5). As in the scalar case, (2.12) implies that when $\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}} > \boldsymbol{0}$ (i.e., the components of $\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}} > \boldsymbol{0}$ are all positive), $\boldsymbol{x}_{\lambda,\delta} = \boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}}$. We define $p_{\mathrm{UC}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$ and $p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$ to be the probability density functions for $\boldsymbol{x}_{\lambda,\delta}^{\mathrm{UC}}$ and $\boldsymbol{x}_{\lambda,\delta}$, respectively. Moreover, we define $\Omega = \{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{x} > \boldsymbol{0}\}$ and its boundary $\mathcal{B} = \{\boldsymbol{x} \geq \boldsymbol{0} \mid x_i = 0 \text{ for some } i\}$. Then by (2.12), the random vector $\boldsymbol{x}_{\lambda,\delta}$ has positive mass, $p_{\mathcal{B}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$, on $\mathcal{B}$ and

$$p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta) = \mathbf{1}_{\Omega}(\boldsymbol{x})\,p_{\mathrm{UC}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta) + \delta_{\mathcal{B}}(\boldsymbol{x})p_{\mathcal{B}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta), \tag{2.14}$$

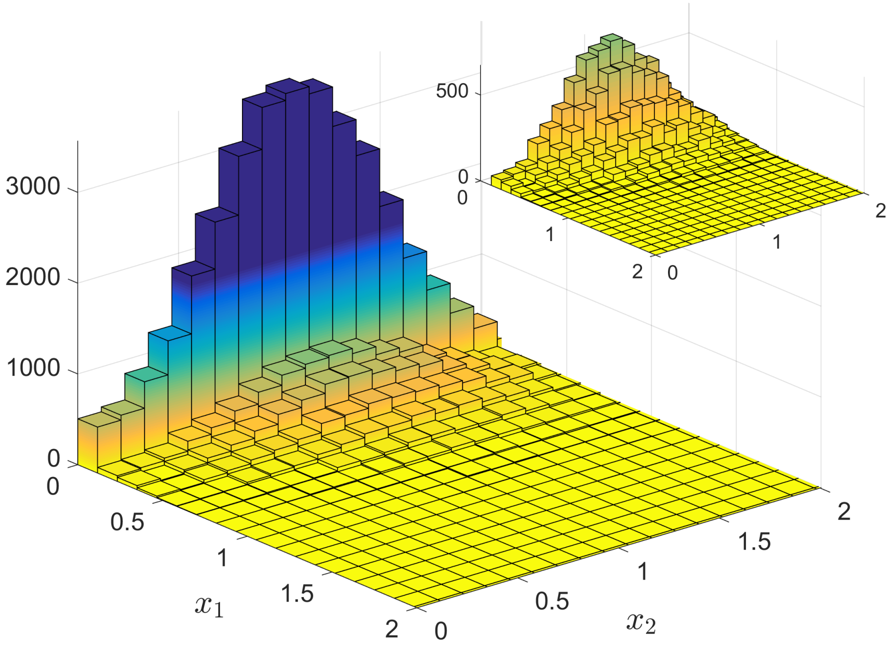

Fig. 3. *The main plot shows a histogram of a simple nonnegativity constrained least squares problem in $\mathbb{R}^2$ whose exact solution is $\boldsymbol{x} = (0,1)^T$. The inset plot shows the histogram when the solutions with the active constraint $x_1 = 0$ are removed.*

where $\mathbf{1}_\Omega$ is the indicator function on Ω and $\delta_\mathcal{B}$ the delta function that equals 1 on $\mathcal{B}$ and 0 otherwise. Unfortunately, we are not able to compute a closed form for $p_\mathcal{B}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$, but we will say much more about $p_{\text{NN}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$ in section 2.3.

It may be possible to obtain further insights into densities (2.11) and (2.14) by viewing them within the transport map framework [20]. In the scalar case, (2.11) can be viewed as the push forward of the unconstrained density $p_{\text{UC}}(x|b,\lambda,\delta)$ through the operator $f(x) = \max(x,0)$. In the multivariate case, (2.14) can be viewed as the push forward of $p_{\text{UC}}(\boldsymbol{x}|\boldsymbol{b},\lambda,\delta)$ through the projection onto $\{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{x} \geq \mathbf{0}\}$ with respect to the norm $\|\boldsymbol{x}\|^2_{\lambda,\delta} \overset{\text{def}}{=} \boldsymbol{x}^T(\lambda\boldsymbol{A}^T\boldsymbol{A} + \delta\boldsymbol{L})\boldsymbol{x}$. We do not pursue this here, but it could lead to a more complete understanding of (2.14).

To illustrate the probability density implicitly defined by (2.9) (and, equivalently–(2.12)–(2.14)), we generated 50000 solutions to a nonnegativity constrained stochastic least squares problem whose exact solution is $\boldsymbol{x} = (0,1)^T$. Figure 3 shows the corresponding histogram, and we see that the constraint $x_1 \geq 0$ is active (i.e., $x_1 = 0$) in many instances—approximately half of the computed solutions. The remaining part of the histogram, for $x_1 > 0$ (which is also shown in the inset plot), is the unnormalized truncated (to x_1, $x_2 > 0$) two-dimensional Gaussian.

Before continuing, we note that our approach is closely related to use of *spike-and-slab priors* for variable selection problems in statistics [16, 25]. The idea behind spike-and-slab-priors is that each coefficient, x_i, of $\boldsymbol{x}$ in the linear model (2.1) has some positive probability of being zero. The spike-and-slab prior is then a mixture of a point mass (Dirac spike) at 0 and a "slab," which is a continuous distribution, e.g., a truncated Gaussian defined on $(0,\infty)$ or a uniform distribution defined on $(0,C)$ for some $C > 0$. A spike-and-slab prior is assumed for each component of x_i of $\boldsymbol{x}$, from which a spike-and-slab prior for $\boldsymbol{x}$ is constructed. This is an intuitive idea, but it is very computationally demanding for large-scale problems, such as are our interest. Our approach corresponds to a spike-and-slab prior that is implemented implicitly through the solution of the optimization problem (2.9) rather than explicitly through the definition of specific prior and hyperprior distributions. It would be interesting to see exactly how our approach fits within the spike-and-slab prior framework, but we do not pursue this here.

2.3. Nonnegative hierarchical Gibbs: Gaussian measurements. We continue with a discussion of (2.9), but first some definitions. We define the *active set* and the *inactive set* for $\boldsymbol{x} \geq \boldsymbol{0}$ by

$$\mathcal{A}(\boldsymbol{x}) = \{i \,|\, x_i = 0\} \qquad \text{and} \qquad \mathcal{I}(\boldsymbol{x}) = \{i \,|\, i \notin \mathcal{A}(\boldsymbol{x})\},$$

respectively, and we note that they consist of complementary subsets of $\{1, \dots, N\}$. Note that if $\mathcal{A}(\boldsymbol{x}) \neq \emptyset$, then $\boldsymbol{x} \in \mathcal{B}$ and $\delta_{\mathcal{B}}(\boldsymbol{x}) = 1$ in (2.14). Finally, we define the matrices $\mathbf{D}_{\mathcal{I}}$ and $\mathbf{D}_{\mathcal{A}}$, which are obtained by removing rows $i \in \mathcal{A}$ and $i \in \mathcal{I}$, respectively, from $\boldsymbol{I}_N$, and corresponding vectors

$$\boldsymbol{x}_{\mathcal{I}} = \mathbf{D}_{\mathcal{I}}\, \boldsymbol{x} \qquad \text{and} \qquad \boldsymbol{x}_{\mathcal{A}} = \mathbf{D}_{\mathcal{A}}\, \boldsymbol{x}.$$

Note that for every $\boldsymbol{x} \in \mathcal{B}$, $\boldsymbol{x}_{\mathcal{I}} > \boldsymbol{0}$ and $\boldsymbol{x}_{\mathcal{A}} = \boldsymbol{0}$.

Returning to (2.9), note that $\boldsymbol{x}_{\lambda,\delta}$ is a random vector, and hence the inactive set $\mathcal{I}(\boldsymbol{x}_{\lambda,\delta})$ is also random. Thus, (2.9) implicitly defines a probability density function $p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I} \,|\, \boldsymbol{b}, \lambda, \delta)$, which we assume is well-defined. In certain restrictive cases, it is possible to analytically compute $p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I} \,|\, \boldsymbol{b}, \lambda, \delta)$ (e.g., when $N = 1$, as was shown above), but in the general case, the calculation is very involved and perhaps even intractable. However, by the laws of conditional probability, we can write

$$p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I} \,|\, \boldsymbol{b}, \lambda, \delta) = p_{\mathrm{NN}}(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda, \delta, \mathcal{I}) p_{\mathrm{NN}}(\mathcal{I} | \boldsymbol{b}, \lambda, \delta), \tag{2.15}$$

which we will study instead of (2.14). We are not able to compute a closed-form expression for $p_{\mathrm{NN}}(\mathcal{I} | \boldsymbol{b}, \lambda, \delta)$, but it is well-defined since $p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I} \,|\, \boldsymbol{b}, \lambda, \delta)$ is assumed to be well-defined. The probability density $p_{\mathrm{NN}}(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda, \delta, \mathcal{I})$, on the other hand, has the form

$$\begin{aligned} p_{\mathrm{NN}}(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda, \delta, \mathcal{I}) &\propto \mathbf{1}_{\Omega}(\boldsymbol{x}) p(\boldsymbol{b} | \boldsymbol{x}, \lambda, \mathcal{I}) p(\boldsymbol{x} | \delta, \mathcal{I}) \\ &\propto \mathbf{1}_{\Omega}(\boldsymbol{x}) \exp\left(-\frac{\lambda}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \frac{\delta}{2} \boldsymbol{x}^T \boldsymbol{L} \boldsymbol{x} \right) \delta_0(\mathbf{D}_{\mathcal{A}} \boldsymbol{x}) && (2.16) \\ &\propto \mathbf{1}_{\Omega_{\mathcal{I}}}(\boldsymbol{x}_{\mathcal{I}}) \exp\left(-\frac{\lambda}{2} \|\boldsymbol{A}_{\mathcal{I}} \boldsymbol{x}_{\mathcal{I}} - \boldsymbol{b}\|_2^2 - \frac{\delta}{2} \boldsymbol{x}_{\mathcal{I}}^T \boldsymbol{L}_{\mathcal{I}} \boldsymbol{x}_{\mathcal{I}} \right) \delta_0(\boldsymbol{x}_{\mathcal{A}}), && (2.17) \end{aligned}$$

where $\boldsymbol{A}_{\mathcal{I}} = \boldsymbol{A}\mathbf{D}_{\mathcal{I}}^T$, $\boldsymbol{L}_{\mathcal{I}} = \mathbf{D}_{\mathcal{I}} \boldsymbol{L} \mathbf{D}_{\mathcal{I}}^T$; $\mathbf{1}_{\Omega_{\mathcal{I}}}$ is the indicator function on $\Omega_{\mathcal{I}} = \{\boldsymbol{x}_{\mathcal{I}} | \boldsymbol{x}_{\mathcal{I}} > \boldsymbol{0}\}$; and $\delta_0(\boldsymbol{x}_{\mathcal{A}})$ implies $\boldsymbol{x}_{\mathcal{A}} = \boldsymbol{0}$ with probability 1.

Both (2.16) and (2.17) are useful ways to express $p_{\mathrm{NN}}(\boldsymbol{x} \,|\, \boldsymbol{b}, \lambda, \delta, \mathcal{I})$. Note that (2.16) shows that $p_{\mathrm{NN}}(\boldsymbol{b} \,|\, \boldsymbol{x}, \lambda, \mathcal{I}) = p(\boldsymbol{b} \,|\, \boldsymbol{x}, \lambda)$, where $p(\boldsymbol{b} \,|\, \boldsymbol{x}, \lambda)$ is defined in (2.2). Moreover, (2.16) and (2.17) give two equivalent expressions for the prior:

$$\begin{aligned} p_{\mathrm{NN}}(\boldsymbol{x} \,|\, \delta, \mathcal{I}) &\propto \exp\left(-\frac{\delta}{2} \boldsymbol{x}^T \boldsymbol{L} \boldsymbol{x} \right) \delta_0(\mathbf{D}_{\mathcal{A}} \boldsymbol{x}) \\ &= \exp\left(-\frac{\delta}{2} \boldsymbol{x}_{\mathcal{I}}^T \boldsymbol{L}_{\mathcal{I}} \boldsymbol{x}_{\mathcal{I}} \right) \delta_0(\boldsymbol{x}_{\mathcal{A}}). \end{aligned}$$

The second equation yields the normalization constant for $p_{\rm NN}(\boldsymbol{x}|\delta, \mathcal{I})$:

$$\int p_{\rm NN}(\boldsymbol{x}|\delta,\mathcal{I})\,d\boldsymbol{x} = \int \delta_0(\boldsymbol{x}_\mathcal{A})\,d\boldsymbol{x}_\mathcal{A} \int \exp\left(-\frac{\delta}{2}\boldsymbol{x}_\mathcal{I}^T \boldsymbol{L}_\mathcal{I}\boldsymbol{x}_\mathcal{I}\right) d\boldsymbol{x}_\mathcal{I}$$

$$= \sqrt{\frac{(2\pi)^{N_\mathcal{I}}}{\det(\delta \boldsymbol{L}_\mathcal{I})}}.$$

Since $\boldsymbol{L}_\mathcal{I} \in \mathbb{R}^{N_\mathcal{I}\times N_\mathcal{I}}$, where $N_\mathcal{I}$ is the number of elements in $\mathcal{I}$, $\det(\delta\boldsymbol{L}_\mathcal{I}) \propto \delta^{N_\mathcal{I}}$, and hence

$$p_{\rm NN}(\boldsymbol{x}\,|\,\delta,\mathcal{I}) \propto \delta^{N_\mathcal{I}/2} \exp\left(-\frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}\right)\delta_0(\mathbf{D}_\mathcal{A}\boldsymbol{x}). \tag{2.18}$$

To complete the hierarchical model, assuming that $\mathcal{I}$ is known, we assume the Gamma hyperpriors (2.6)–(2.7) for λ and δ and obtain, via Bayes' law,

$$\begin{aligned} p_{\rm NN}(\boldsymbol{x},\lambda,\delta\,|\,\boldsymbol{b},\mathcal{I}) &\propto \mathbf{1}_\Omega(\boldsymbol{x})p(\boldsymbol{b}\,|\,\boldsymbol{x},\lambda,\mathcal{I})\,p_{\rm NN}(\boldsymbol{x}\,|\,\delta,\mathcal{I})\,p(\lambda)\,p(\delta) \\ &= \lambda^{M/2+\alpha_\lambda-1}\,\delta^{N_\mathcal{I}/2+\alpha_\delta-1}\times\mathbf{1}_\Omega(\boldsymbol{x})\times\delta_0(\mathbf{D}_\mathcal{A}\boldsymbol{x}) \\ &\quad\times\exp\left(-\frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2-\frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}-\beta_\lambda\lambda-\beta_\delta\delta\right), \quad \lambda,\delta>0, \end{aligned} \tag{2.19}$$

which is well-defined. Finally, using the laws of conditional probability, we have

$$p_{\rm NN}(\boldsymbol{x},\mathcal{I},\lambda,\delta|\boldsymbol{b}) = p_{\rm NN}(\boldsymbol{x},\lambda,\delta\,|\,\boldsymbol{b},\mathcal{I})p_{\rm NN}(\mathcal{I}|\boldsymbol{b}), \tag{2.20}$$

which is well-defined provided $p_{\rm NN}(\mathcal{I}|\boldsymbol{b})$ is a well-defined probability mass function.

From (2.19), we obtain the conditional density $p_{\rm NN}(\lambda,\delta\,|\,\boldsymbol{b},\boldsymbol{x},\mathcal{I})$, and from (2.9), we obtain (implicitly) $p_{\rm NN}(\boldsymbol{x},\mathcal{I}\,|\,\boldsymbol{b},\lambda,\delta)$. These two conditional densities are used to define the nonnegative hierarchical Gibbs sampler (NHGS), which originally appeared in [4]. The target distribution for NHGS is given by $p_{\rm NN}(\boldsymbol{x},\mathcal{I},\lambda,\delta|\boldsymbol{b})$ defined in (2.20).

Nonnegative Hierarchical Gibbs Sampler (NHGS)

```
Initialization: Choose $(\boldsymbol{x}^0, \mathcal{I}_0, \lambda_0, \delta_0)$.
For $k = 1, 2, \ldots$
    Compute $(\lambda_k, \delta_k) \sim p_{\rm NN}(\lambda, \delta\,|\,\boldsymbol{b}, \boldsymbol{x}^{k-1}, \mathcal{I}_{k-1})$ as follows:
        $\lambda_k \sim \Gamma\left(M/2 + \alpha_\lambda, \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^{k-1} - \boldsymbol{b}\|_2^2 + \beta_\lambda\right)$,
        $\delta_k \sim \Gamma\left(N_{\mathcal{I}_{k-1}}/2 + \alpha_\delta, \frac{1}{2}(\boldsymbol{x}^{k-1})^T\boldsymbol{L}\boldsymbol{x}^{k-1} + \beta_\delta\right)$.
    Compute $(\boldsymbol{x}^k, \mathcal{I}_k) \sim p_{\rm NN}(\boldsymbol{x}, \mathcal{I}\,|\,\boldsymbol{b}, \lambda_k, \delta_k)$ using (2.9)
        with $(\lambda, \delta) = (\lambda_k, \delta_k)$.
End.
```

Remarks. An obvious question is, what are the convergence properties of NHGS? First, note that NHGS is a two-stage Gibbs sampler [26, Chapter 9] and that we have defined the conditional densities $p_{\rm NN}(\lambda,\delta|\boldsymbol{b},\boldsymbol{x},\mathcal{I})$ and (implicitly through (2.9)) $p_{\rm NN}(\boldsymbol{x},\mathcal{I}|\boldsymbol{b},\lambda,\delta)$. Unfortunately, well-defined conditional densities do not guarantee a well-defined joint density [26, section 10.4.3], but we have defined the joint density (2.20) and assumed that it is well-defined. We assume, further, that $p_{\rm NN}(\boldsymbol{x},\mathcal{I},\lambda,\delta|\boldsymbol{b})$ satisfies the requirements for the convergence of a two-stage Gibbs sampler, in which case it will be the target distribution of NHGS. Moreover, two-stage Gibbs samplers have a number of special properties, one of which is that the individual chains

$\{(\lambda_k, \delta_k)\}_{k=1}^{\infty}$ and $\{(\boldsymbol{x}_k, \mathcal{I}_k)\}_{k=1}^{\infty}$ are reversible [26, Lemma 9.11], and while the joint chain $\{(\boldsymbol{x}_k, \mathcal{I}_k, \lambda_k, \delta_k)\}_{k=1}^{\infty}$ is not necessarily reversible, it does satisfy an "interleaving property" as described in [26, Lemma 9.11]. One important consequence of NHGS being a two-stage Gibbs sampler is that it suffices to monitor the convergence of the (λ, δ)-chain to determine the convergence of the joint $(\boldsymbol{x}, \mathcal{I}, \lambda, \delta)$-chain, which is a significant advantage given that $(\boldsymbol{x}, \mathcal{I})$ is high-dimensional.

Another question is, what is special about linear inverse problems; i.e., can NHGS be used for nonlinear inverse problems? In theory, one can simply replace the linear function $\boldsymbol{A}\boldsymbol{x}$ by a nonlinear function $\boldsymbol{A}(\boldsymbol{x})$ in (2.9) and then define $p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I}|\boldsymbol{b}, \lambda, \delta)$ as in the linear case, though this requires that (2.9) has a unique solution for every $\hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b}, \lambda^{-1}\boldsymbol{I}_M)$ and $\hat{\mathbf{c}} \sim \mathcal{N}(\mathbf{0}, \delta^{-1}\boldsymbol{I}_N)$ and that it can be solved efficiently, both of which are not guaranteed. In the linear case, (2.17) shows that $p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b}, \lambda, \delta, \mathcal{I})$ decouples into a truncated Gaussian distribution on $\boldsymbol{x}_{\mathcal{I}}$ and a Dirac delta function on $\boldsymbol{x}_{\mathcal{A}}$. As a consequence, $p_{\mathrm{NN}}(\boldsymbol{x}, \lambda, \delta|\boldsymbol{b}, \mathcal{I})$ defined in (2.19) corresponds to a standard Gaussian-Gamma hierarchical model on $\boldsymbol{x}_{\mathcal{I}}$. Moreover, whereas (2.9) yields exact samples from $p_{\mathrm{NN}}(\boldsymbol{x}|\boldsymbol{b}, \lambda, \delta, \mathcal{I})$, this is not true in the nonlinear case. One could correct this by using randomize-then-optimize [5] as a proposal within Metropolis-Hastings, but we do not pursue that here.

3. Poisson measurement error.

In this section, we assume, instead, that the data $\boldsymbol{b}$ arise from a Poisson distribution, which is the case in both astronomical imaging and positron emission tomography [24, 28]. Specifically, we assume

$$\boldsymbol{b} = \mathrm{Poiss}(\boldsymbol{A}\boldsymbol{x} + \mathbf{g}), \tag{3.1}$$

where $\mathbf{g}$ is the $N \times 1$ vector of *known* background counts. Then the probability density function for the data is given by

$$p(\boldsymbol{b}|\boldsymbol{x}) = \prod_{i=1}^{N} \frac{([\boldsymbol{A}\boldsymbol{x}]_i + g_i)^{b_i} \exp[-([\boldsymbol{A}\boldsymbol{x}]_i + g_i)]}{b_i!}, \quad \boldsymbol{x} \geq \mathbf{0}. \tag{3.2}$$

Assuming the prior $p(\boldsymbol{x}|\delta)$ defined by (2.3), the posterior density $p(\boldsymbol{x}|\boldsymbol{b}, \delta)$ has the form

$$p(\boldsymbol{x}|\boldsymbol{b}, \delta) \propto \exp\left(-\sum_{i=1}^{n} \{[\boldsymbol{A}\boldsymbol{x}]_i + g_i - b_i \ln([\boldsymbol{A}\boldsymbol{x}]_i + g_i)\} + \frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x}\right). \tag{3.3}$$

To obtain a nonnegativity constrained sampler in the Poisson measurement case, we make intuitive changes to the stochastic, nonnegativity constrained least squares problem (2.9). First, we replace the Gaussian negative log-likelihood by the Poisson negative log-likelihood, and, second, we replace $\hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b}, \lambda^{-1})$ by $\hat{\boldsymbol{b}} \sim \mathrm{Poiss}(\boldsymbol{b})$. The resulting nonnegativity constrained, stochastic, convex optimization problem is given by

$$\begin{gathered}\boldsymbol{x}_\delta = \arg\min_{\boldsymbol{x} \geq \mathbf{0}} \sum_{i=1}^{n} \{[\boldsymbol{A}\boldsymbol{x}]_i + g_i - \hat{b}_i \ln([\boldsymbol{A}\boldsymbol{x}]_i + g_i)\} + \frac{\delta}{2}\|\boldsymbol{L}^{1/2}\boldsymbol{x} - \hat{\mathbf{c}}\|^2, \\ \text{where } \hat{\boldsymbol{b}} \sim \mathrm{Poiss}(\boldsymbol{b}) \text{ and } \hat{\mathbf{c}} \sim \mathcal{N}(\mathbf{0}, \delta^{-1}\boldsymbol{I}).\end{gathered} \tag{3.4}$$

To solve (3.4), we use the active set method gradient project reduced Newton, which was developed in [6] for precisely such problems.

3.1. A Bayesian hierarchical model and Gibbs sampler.

Just as was the case for (2.9), $\boldsymbol{x}_\delta$ defined by (3.4) is a random vector with associated random inactive set $\mathcal{I}(\boldsymbol{x}_\delta)$. We denote the resulting probability density for $(\boldsymbol{x}, \mathcal{I})$ by $p_{\rm NN}(\boldsymbol{x}, \mathcal{I}|\boldsymbol{b}, \delta)$. As in the Gaussian measurement case, we cannot compute an analytic expression for $p_{\rm NN}(\boldsymbol{x}, \mathcal{I}|\boldsymbol{b}, \delta)$, but by the laws of conditional probability, we know that

$$p_{\rm NN}(\boldsymbol{x} \mid \boldsymbol{b}, \mathcal{I}, \delta) \propto p_{\rm NN}(\boldsymbol{b}|\boldsymbol{x}, \mathcal{I}) p_{\rm NN}(\boldsymbol{x}|\mathcal{I}, \delta),$$

where $p_{\rm NN}(\boldsymbol{b}|\boldsymbol{x}, \mathcal{I}) = p(\boldsymbol{b}|\boldsymbol{x})$ and $p_{\rm NN}(\boldsymbol{x}|\delta, \mathcal{I})$ is given by (2.18).

As in the Gaussian measurement case, we assume the hyperprior $p(\delta)$ defined by (2.7), and following similar arguments, we obtain, through Bayes' law, the conditional density

$$\begin{aligned} p_{\rm NN}(\boldsymbol{x}, \delta|\boldsymbol{b}, \mathcal{I}) &\propto p_{\rm NN}(\boldsymbol{x} \mid \boldsymbol{b}, \mathcal{I}, \delta) p(\delta) \\ &\propto \delta^{N_\mathcal{I}/2+\alpha_\delta-1} \times \delta_0(\mathbf{D}_\mathcal{A}\boldsymbol{x}) \\ &\quad \times \exp\left(-\sum_{i=1}^n \{[\boldsymbol{A}\boldsymbol{x}]_i + g_i - b_i \ln([\boldsymbol{A}\boldsymbol{x}]_i + g_i)\} - \frac{\delta}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x} - \beta_\delta\delta \right), \quad \delta > 0. \end{aligned} \tag{3.5}$$

The conditional density $p_{\rm NN}(\delta|\boldsymbol{b}, \boldsymbol{x}, \mathcal{I})$ is then given by

$$p_{\rm NN}(\delta|\boldsymbol{b}, \boldsymbol{x}, \mathcal{I}) = \delta^{N_\mathcal{I}/2+\alpha_\delta-1} \exp\left(\left[-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{L}\boldsymbol{x} - \beta_\delta \right] \delta \right). \tag{3.6}$$

We use this, together with $p_{\rm NN}(\boldsymbol{x}, \mathcal{I}|\boldsymbol{b}, \delta)$ defined implicitly by (3.4), to define the following nonnegative Gibbs sampler for Poisson distributed measurements.

Poisson Hierarchical Gibbs Sampler

```
Initialization: Choose $(\boldsymbol{x}^0, \mathcal{I}_0, \delta_0)$.
For $k = 1, 2, \ldots$
    Compute $\delta_k \sim p_{\rm NN}(\delta \mid \boldsymbol{b}, \boldsymbol{x}^{k-1}, \mathcal{I}_{k-1})$ as follows:
        $\delta_k \sim \Gamma\left(N_{\mathcal{I}_{k-1}}/2 + \alpha_\delta, \frac{1}{2}(\boldsymbol{x}^{k-1})^T\boldsymbol{L}\boldsymbol{x}^{k-1} + \beta_\delta\right)$.
    Compute $(\boldsymbol{x}^k, \mathcal{I}_k) \sim p_{\rm NN}(\boldsymbol{x}, \mathcal{I} \mid \boldsymbol{b}, \delta_k)$ using (3.4) with $\delta = \delta_k$.
End.
```

In contrast to the Gaussian measurement case, where the regularization parameter is the ratio (δ/λ), in this case, the δ is precisely the regularization parameter.

4. Numerical examples.

In this section, we implement both nonnegative hierarchical Gibbs and Poisson hierarchical Gibbs on synthetic test cases from image deblurring, in both one and two dimensions, and on a test case from positron emission tomography. We also illustrate the dimension dependence of both of these MCMC methods.

4.1. One-dimensional image deblurring.

We begin with a one-dimensional image deblurring example. The measurement model is of the form (2.1) for Gaussian measurements and (3.1) for Poisson measurements, with the discretized model $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ obtained via midpoint quadrature [2] applied to the convolution equation

$$b(s) = \int_0^1 A(s - s')x(s')ds'$$

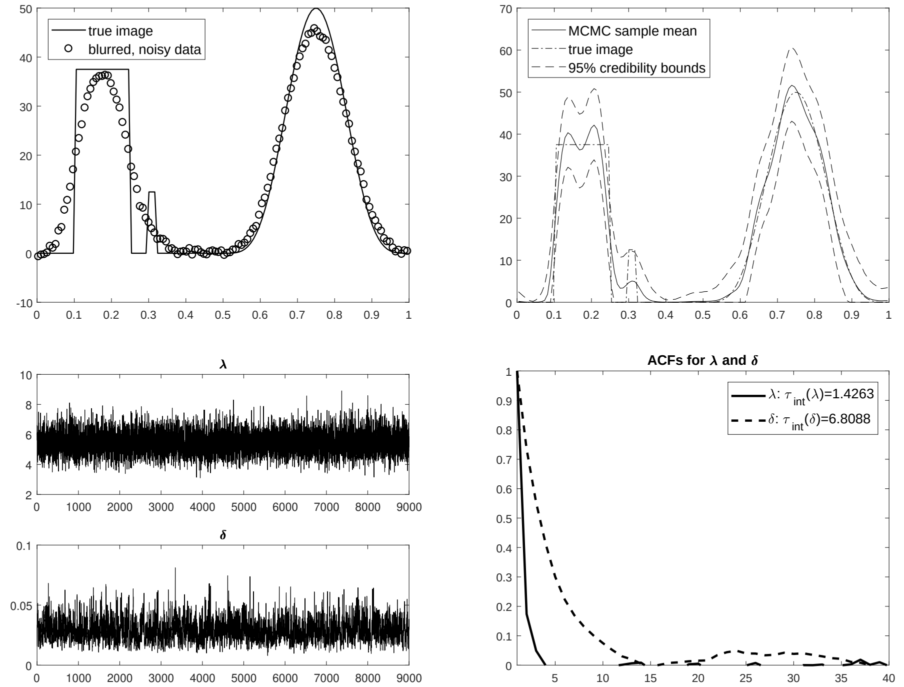

FIG. 4. *UQ using nonnegative hierarchical Gibbs for Gaussian measurements. In the upper left is a plot of a true image $\boldsymbol{x}$ and blurred, noisy data $\boldsymbol{b}$ from the one-dimensional image deblurring test problem for $N = 128$. In the upper right is a plot of the sample mean and 95% credibility bounds for the $\boldsymbol{x}$-samples. In the lower left are chain plots for the λ- and δ-samples, while in the lower right are plots of the corresponding ACFs, and the IACTs are given in the legend.*

with a Gaussian convolution kernel $A(s) = \exp(-s^2/(2\gamma^2))/\sqrt{\pi\gamma^2}$, $\gamma > 0$. Then $\boldsymbol{A}$ has the form

$$[\boldsymbol{A}]_{ij} = h \exp\left(-((i-j)h)^2/(2\gamma^2)\right)/\sqrt{\pi\gamma^2}, \quad 1 \le i, j \le N, \tag{4.1}$$

where $h = 1/N$ with N the number of grid points in $[0, 1]$. We use $N = 128$ in our tests, and the resulting $\boldsymbol{A}$ is invertible but has condition number on the order of 10^{16}, resulting in a severely ill conditioned problem. The true image and blurred noisy data are plotted in the upper left of Figure 4. In order to avoid committing an “inverse crime” [19], we generate our measurements using the finer mesh corresponding to $N = 256$.

In the Gaussian measurement case, we implement NHGS, taking 10000 samples and discarding the first 1000 as burn-in. The sample mean and 95% credibility bounds for $\boldsymbol{x}$ are plotted in the upper right in Figure 4. The individual chain plots for the λ- and δ-chains are plotted in the lower left in Figure 4. To assess convergence of the MCMC chain, we computed Geweke p-values [13] for the λ- and δ-chains, both of which were approximately 0.99, providing strong evidence of convergence. As can be seen by seen in the legend of the autocorrelation plot in the lower right in Figure 4, the integrated autocorrelation time (IACT) for the resulting MCMC chain is approximately 7, which means that the correlated 9000-sample chain contains roughly 1200 independent samples. The IACT is defined from the autocorrelation function

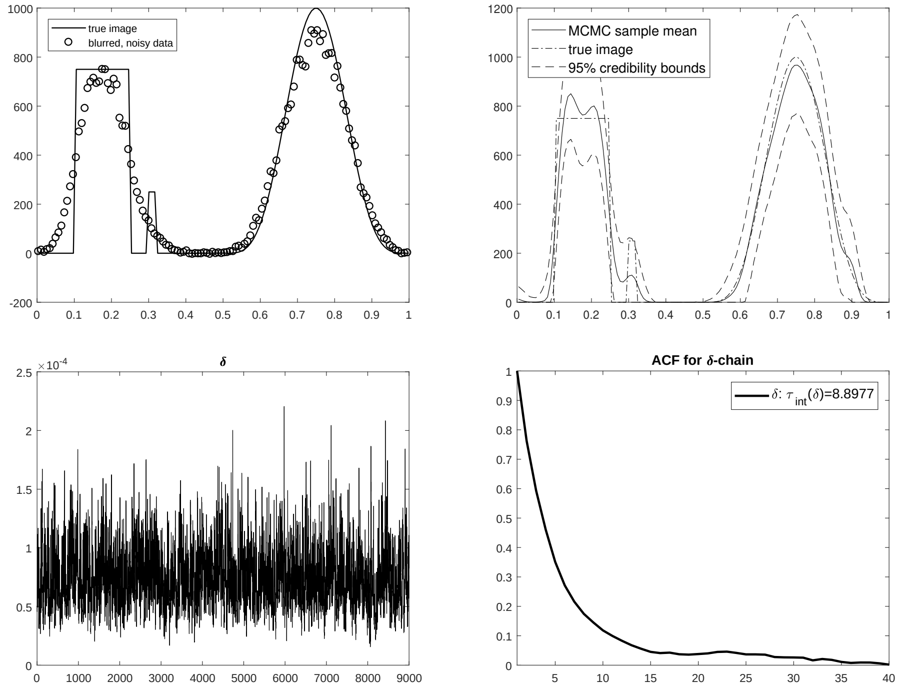

FIG. 5. *UQ using Poisson hierarchical Gibbs. In the upper left is a plot of a true image $\boldsymbol{x}$ and blurred, noisy data $\boldsymbol{b}$ from the one-dimensional image deblurring test problem for $N = 128$. In the upper right is a plot of the sample mean and 95% credibility bounds for the $\boldsymbol{x}$-samples. In the lower left are chain plots for the δ-samples, while in the lower right is a plot of the corresponding ACF, and the IACT is also given.*

(ACF) of a Markov chain. For $\{\delta_k\}_{k=1}^K$, the ACF is defined as

$$\hat{\rho}(j) = C(j)/C(0), \quad \text{where} \quad C(j) = \frac{1}{K-j}\sum_{k=1}^{K-j}(\delta_k - \bar{\delta})(\delta_{k+|j|} - \bar{\delta}), \tag{4.2}$$

where $\bar{\delta} = \frac{1}{K}\sum_{k=1}^K \delta_k$. The faster $\hat{\rho}(j)$ decays to zero, the less correlated is the δ-chain. The IACT for $\{\delta_k\}_{k=1}^K$, denoted $\tau_{\text{int}}(\delta)$, is obtained from the ACF (see [2] for details); specifically, the faster (slower) $\hat{\rho}(j)$ decays to zero, the smaller (larger) will be the IACT. The IACT can be viewed as the minimal distance between statistically independent elements of the δ-chain; i.e., δ_k and $\delta_{k\pm j}$ will be independent provided that $j \geq \tau_{\text{int}}(\delta)$.

In the Poisson measurement case, we apply Poisson hierarchical Gibbs, again taking 10000 samples and discarding the first 1000 as burn-in. The true image and blurred noisy data are plotted in the upper left. The sample mean and 95% credibility bounds are plotted in the upper right in Figure 5. The individual chain plot for the δ-chain is shown in the lower left, and the ACF (with IACT$\approx$ 9) is plotted in the lower right. To assess convergence of the MCMC chain, we computed Geweke p-values [13] for the δ-chain, obtaining a value of 0.93, which provides strong evidence of convergence. Thus, the correlated 9000-sample chain contains roughly 1000 independent samples, suggesting that the estimates of the mean and variance should be stable and have converged.

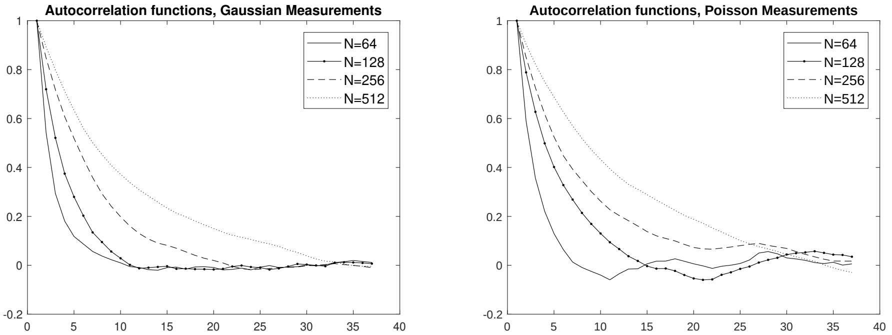

Fig. 6. *Plots of the autocorrelation functions for δ-chains generated by nonnegative hierarchical Gibbs (left) and Poisson hierarchical Gibbs (right), showing that both exhibit dimension-dependent behavior. Test case: one-dimensional image deblurring with $N = 64$, 128, 256, and 512 grid points.*

4.1.1. Degeneracy in the hierarchical Gibbs samplers. When the discrete model $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ in (2.1) arises as the discretization of an integral equation of the form $b(s) = \int_D a(s,t)\, x(t)\, dt$, the parameter vector $\boldsymbol{x}$ is the discretization of a function x defined on the computational domain D. Thus, it is natural to ask if the performance of the above Gibbs samplers is dependent upon N, especially as $N \to \infty$. In [1, Theorem 3.4], it is shown that for hierarchical Gibbs, under reasonable assumptions, the expected value of the step $\Delta_k = \delta_{k+1} - \delta_k$ at step k in the δ-chain, given fixed δ_k, scales like $2/N$; specifically,

$$\frac{N}{2}\mathbb{E}\left[\Delta_k \,|\, \delta_k\right] = (\alpha_\delta + 1)\delta_k - f_N(\delta_k; \boldsymbol{b})\delta_k^2 + \mathcal{O}(N^{-1/2}),$$

where $\mathbb{E}$ denotes expectation and $f_N(\delta_k; \boldsymbol{b})$ is bounded uniformly as $N \to \infty$. The variance of the step also scales like $2/N$; specifically, for any $\delta > 0$,

$$\frac{N}{2}\mathrm{Var}\left[\Delta_k \,|\, \delta_k\right] = 2\delta_k^2 + \mathcal{O}(N^{-1/2}).$$

Thus, the δ-chain becomes increasingly correlated as $N \to \infty$.

It is desirable to develop and use MCMC algorithms whose behavior is independent of the discretization dimension N. The results from the previous paragraph, originally in [1], show that the behavior of hierarchical Gibbs is dimension *dependent.* Since we are using the same prior in the nonnegative and Poisson hierarchical Gibbs algorithms, we expect that the correlation in the resulting δ-chains will also be dependent upon N. We verify this numerically in Figure 6, which displays the ACFs for the δ-chains generated by the nonnegative (left) and Poisson (right) hierarchical Gibbs algorithms applied to the one-dimensional image deblurring problem defined above for $N = 64$, 128, 256, and 512. It is clear from Figure 6 that the correlation in the δ-chain increases as N increases.

Several alternatives to hierarchical Gibbs are presented in [2] that have behavior that is independent of N. Each of these methods makes use of the marginal density $p(\lambda, \delta \,|\, \boldsymbol{b}) = \int p(\boldsymbol{x}, \lambda, \delta \,|\, \boldsymbol{b})\, d\boldsymbol{x}$. In the nonnegativity constrained case, we have implemented analogous algorithms that make use of the marginal density, $p_{\mathrm{NN}}(\mathcal{I}, \lambda, \delta \,|\, \boldsymbol{b}) = \int p_{\mathrm{NN}}(\boldsymbol{x}, \mathcal{I}, \lambda, \delta \,|\, \boldsymbol{b})\, d\boldsymbol{x}$. However, because $\mathcal{I}$ depends upon $\boldsymbol{x}$, the performance of the resulting algorithms remains dependent on N. To obtain a dimension-independent algorithm, we therefore need to "fully marginalize," i.e., use instead $p_{\rm NN}(\lambda, \delta \,|\, \boldsymbol{b}) = \int p_{\rm NN}(\boldsymbol{x}, \mathcal{I}, \lambda, \delta \,|\, \boldsymbol{b})\, d\boldsymbol{x} d\mathcal{I}$. But we cannot compute this analytically because we do not have an analytic expression for $p_{\rm NN}(\boldsymbol{x}, \mathcal{I}, \lambda, \delta \,|\, \boldsymbol{b})$.

An alternative approach that would yield a dimension-independent constrained sampler was suggested by one of the referees. The approach is in the spirit of [11], and although we do not implement it, we provide a brief description. First, rather than expressing the unknown function $x(t)$ (of which $\boldsymbol{x}$ is a discretization) in terms of the pixel (pointwise) basis, it is expressed in terms of a strictly positive wavelet basis $\{w_k(t)\}_{k=1}^{\infty}$:

$$x(t) = \sum_{i=1}^{\infty} \chi_k w_k(t).$$

Since $x(t)$ is identified by $\{\chi_k\}_{k=1}^{\infty}$, inference is performed on the χ_k's rather than on the point values of $x(t)$. Moreover, a function space Gaussian prior (see, e.g., [29]) for x can be constructed using the wavelet basis, with a covariance matrix chosen so that its eigenvalues decay sufficiently fast that the problem will have finite effective dimension. The NHGS algorithm can then be applied to the problem of estimating the coefficients $\{\chi_k\}_{k=1}^{\infty}$.

4.2. Two-dimensional image deblurring.

In the two-dimensional deblurring case, the model has the form

$$b(s,t) = \int_0^1 \int_0^1 A(s-s', t-t') x(s',t') ds' dt'.$$

For our tests, we choose a Gaussian convolution kernel A and discretize using midpoint quadrature on a 128×128 uniform computational grid over [0,1]×[0,1]. Moreover, we assume periodic boundary conditions so that $\boldsymbol{A}$ is a $128^2 \times 128^2$-block circulant with circulant blocks matrix and hence is diagonalizable by the two-dimensional discrete Fourier transform [2]. We generate data $\boldsymbol{b}$ using both (2.1) and (3.1) with background $\mathbf{g} = 10 \cdot \mathbf{1}$, and in order to avoid committing an inverse crime [19], we generate our measurements using a finer mesh, specifically, a 256×256 uniform computational grid over [0,1]×[0,1].

In the Gaussian measurement case, we apply NHGS, taking 10000 samples and removing the first 1000 as burn-in. The results are shown in Figure 7: In the upper left is a plot of the blurred, noisy data; in the upper right is a plot of the mean of the $\boldsymbol{x}$-chain; and in the lower left is a plot of the *width* pixelwise 95% credibility bounds, all computed from the $\boldsymbol{x}$-chain. And finally, in the lower right is a plot of two histograms created from the λ- and δ-chains. To assess convergence of the MCMC chain, we computed Geweke p-values [13] for the λ- and δ-chains, both of which were approximately 0.99, providing strong evidence of convergence. The IACT for the λ- and δ-chains was approximately 1.3 and 11.5, respectively. Thus, the correlated 9000-sample MCMC chain contains roughly 780 independent samples, suggesting that the estimates of the mean and variance should be stable and have converged.

In the Poisson measurement case, we apply NHGS and again compute a chain of length 10000, removing the first 1000 samples as burn-in. The results are shown in Figure 8: In the upper left is a plot of the blurred, noisy data; in the upper right is a plot of the mean of the $\boldsymbol{x}$-chain; and in the lower left is a plot of the *width* pixelwise 95% credibility bounds, all computed from the $\boldsymbol{x}$-chain. And finally, in the lower right is a plot of the histogram created from the δ-chain. The IACT for the δ-chain is approximately 7.5. Thus, the correlated 9000-sample MCMC chain contains roughly

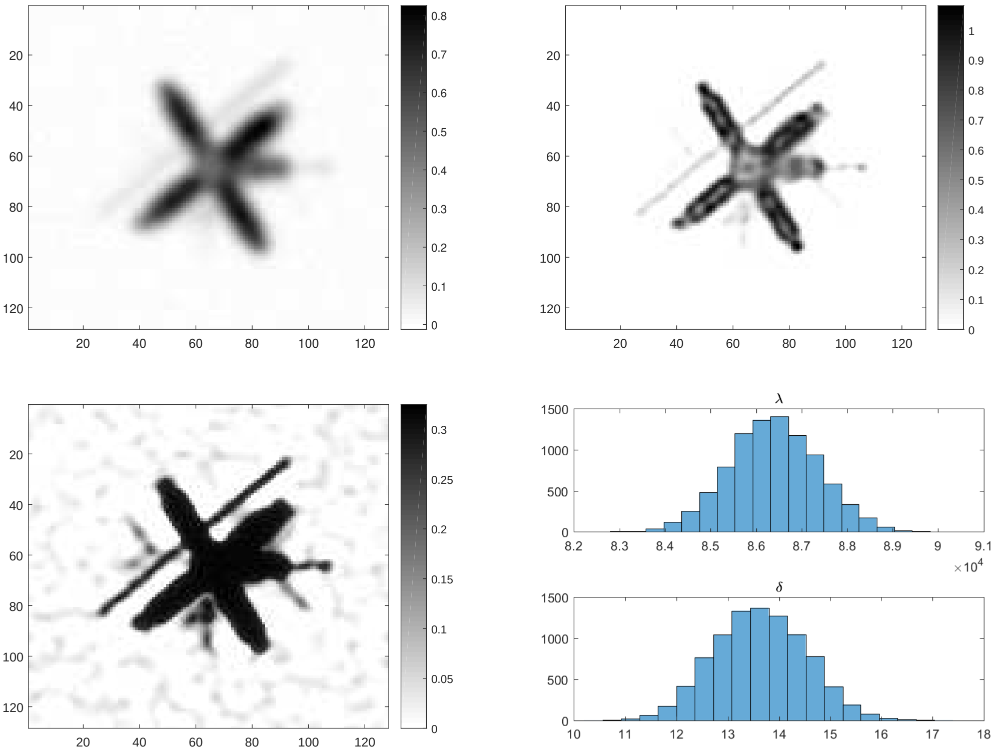

FIG. 7. *Two-dimensional deblurring with Gaussian measurements. In the upper left is a plot of the blurred noisy data. In the upper right is a plot of the mean of the $\boldsymbol{x}$-chain. In the lower left is a plot of the width of the pixelwise* 95% *credibility bands. In the lower right is a histogram of the elements of the* λ*- and* δ*-chains.*

1200 independent samples, suggesting that the estimates of the mean and variance should be stable and have converged.

4.3. Positron emission tomography.

In positron emission tomography (PET), a radioactive tracer element is injected into a body and exhibits radioactive decay, resulting in photon emission. The emitted photons that leave the body are recorded by a photon detector, which also determines the line of response $L \subset \mathbb{R}^2$. In the two-dimensional case, each line L is uniquely determined by (i) the angle, ω, it makes with an axis (e.g., the vertical) and (ii) its perpendicular distance, y, from the origin. Thus, we denote a given line by $L(\omega, y)$ and parameterize it by a scalar $s \in [0, S]$ so that $L(\omega, y) = \{z(s) \mid 0 \le s \le S\}$, where $z(0)$ and $z(s)$ correspond to the positions of the two detectors connected by $L(\omega, y)$.

PET data $b(\omega, y)$ correspond to the number of detected photon pairs along $L(\omega, y)$. The model relating the tracer density x to the data is given by

$$b(\omega, y) = \int_{L(\omega,y)} A_{\omega,y}(z(s))x(z(s))ds,$$

where the impulse response function $A_{\omega,y}(z(r))$ can be viewed as the probability that an emission event located at $z(r)$ along $L(\omega, y)$ is recorded by the detector system. A

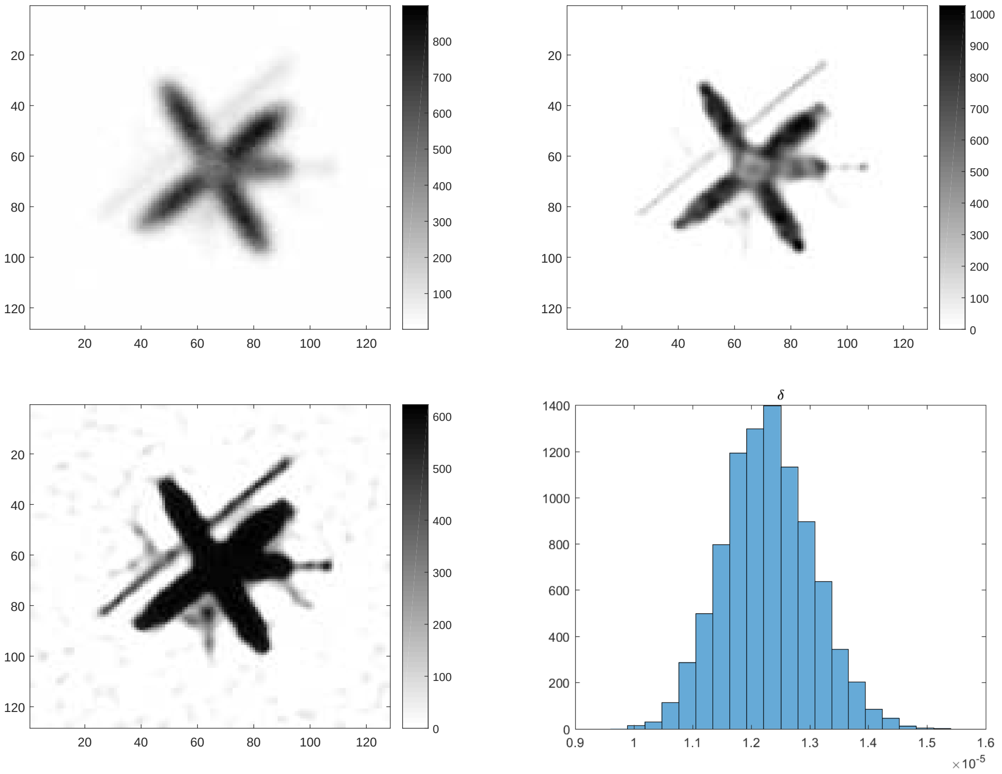

FIG. 8. *Two-dimensional deblurring with Poisson measurements. In the upper left is a plot of the blurred noisy data. In the upper right is a plot of the mean of the* $\boldsymbol{x}$*-chain. In the lower left is a plot of the width of the pixelwise* 95% *credibility bands. In the lower right is a histogram of the elements of the* δ*-chain.*

pair of photons are emitted at a location $z(r)$ along $L(\omega, y)$ with detectors located at $z(0)$ and $z(S)$. In this case, the impulse response is the product of probabilities,

$$A_{\omega,y}(z(r)) = e^{-\int_0^r \mu(z(t))\,dt} e^{-\int_r^S \mu(z(t))\,dt} = e^{-\int_{L(\omega,y)} \mu(z(t))\,dt},$$

which does not depend on r and $\mu(z)$ (the object density). Hence, we have the somewhat simpler mathematical model

$$b(\omega, y) = e^{-\int_{L(\omega,y)} \mu(z(t))\,dt} \int_{L(\omega,y)} x(z(s))ds. \tag{4.3}$$

In fact, dividing both sides of (4.3) by $e^{-\int_{L(\omega,y)} \mu(z(t))\,dt}$ yields the Radon transform model, which is what is solved in the computed tomography inverse problem.

After discretization, (4.3) can be written as a system of linear equations of the form $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$. The discretization occurs both in the spatial domain, where μ and x are defined, and in the Radon transform ((ω, y)) domain, where the data b are defined. We use a uniform $n \times n$ spatial grid and a uniform grid for the transform domain with n angles and n sensors. In our experiments, $n = 100$ so that $\boldsymbol{A}$ has size 10000×10000.

Since the data $\boldsymbol{b}$ consists of photon counts, a Poisson noise model of the form (3.1) is used [15, 24]. We use the Shepp–Logan phantom generated using MATLAB's

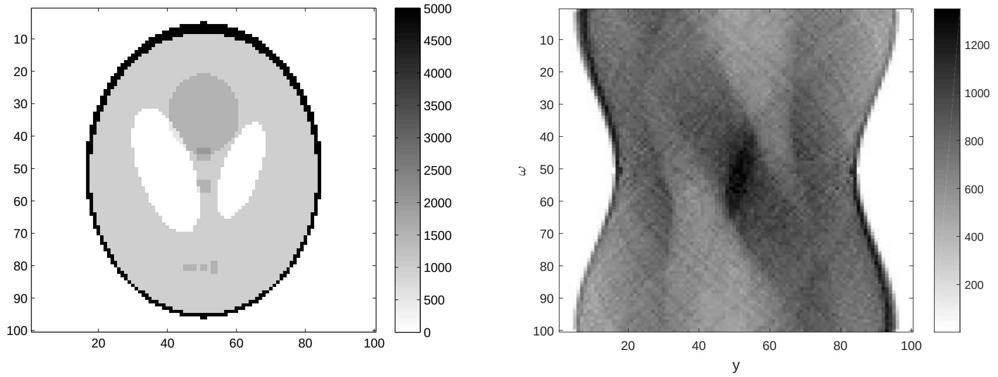

FIG. 9. *On the left is the true image, and on the right are the PET data.*

`phantom` function for our true tracer density. We take $\mu = 0$, which is standard for PET numerical experiments [24]. The true tracer density $\boldsymbol{x}$ is then scaled so that the signal-to-noise ratio is 28. Both the data, which were generated using (3.1) with $\mathbf{g} = 10 \cdot \mathbf{1}$, and the true tracer density in the PET case are shown in Figure 9.

In this implementation, we computed five parallel NHGS chains, each of length 2000, for a total of 10000 samples. We removed the first 200 samples from each chain as burn-in. To determine convergence, we use the $\hat{R}$ estimator defined in [12], which compares convergence to the mean both within each chain and between chains. In this run, $\hat{R} = 1.002$, strongly indicating convergence; hence, accurate estimates of the mean and variance of the parameters can be expected. We used the 9000 samples that remained after burn-in for our analysis, presenting the results in Figure 10. In the upper left is a plot of the reconstructed tracer density obtained by computing the mean of the $\boldsymbol{x}$-samples. The result seems to be visually superior to the MAP estimator $\boldsymbol{x}_\delta$ defined by (3.4), where δ was taken to be the mean of the δ-chain, which plotted in the upper right. We also plot the pixelwise standard deviation image and a histogram for the δ samples, both of which can be found on the bottom in Figure 10. Note that as in the image deblurring example, the variance is higher within the support of the object.

5. Conclusions. We have drawn a distinction between positivity and nonnegativity constraints for probability distribution arising in inverse problems. For us, nonnegativity in a probability distribution means positive mass at the boundary of the set $\{\boldsymbol{x} \in \mathbb{R}^N \,|\, \boldsymbol{x} \geq \mathbf{0}\}$. Nonnegative probability densities are desirable for problems in which the unknown parameters—for us the components of $\boldsymbol{x}$—have a positive probability of being zero. Spike-and-slab priors can be used within a Bayesian framework to obtain nonnegative posterior densities, but they are very computationally demanding for large-scale problems. We have shown how to implement nonnegative probability models for inverse problems through the use of nonnegativity constrained optimization. State-of-the-art numerical optimization methods make our approach computationally efficient even for large-scale problems. We consider both Gaussian and Poisson distributed measurement models, assume Gaussian priors, and then construct a related nonnegative posterior probability density function by defining a stochastic, nonnegativity constrained optimization problem. We then embed the resulting nonnegative samplers within a hierarchical Gibbs sampler to obtain two nonnegativity constrained MCMC methods. The paper ends with numerical tests,

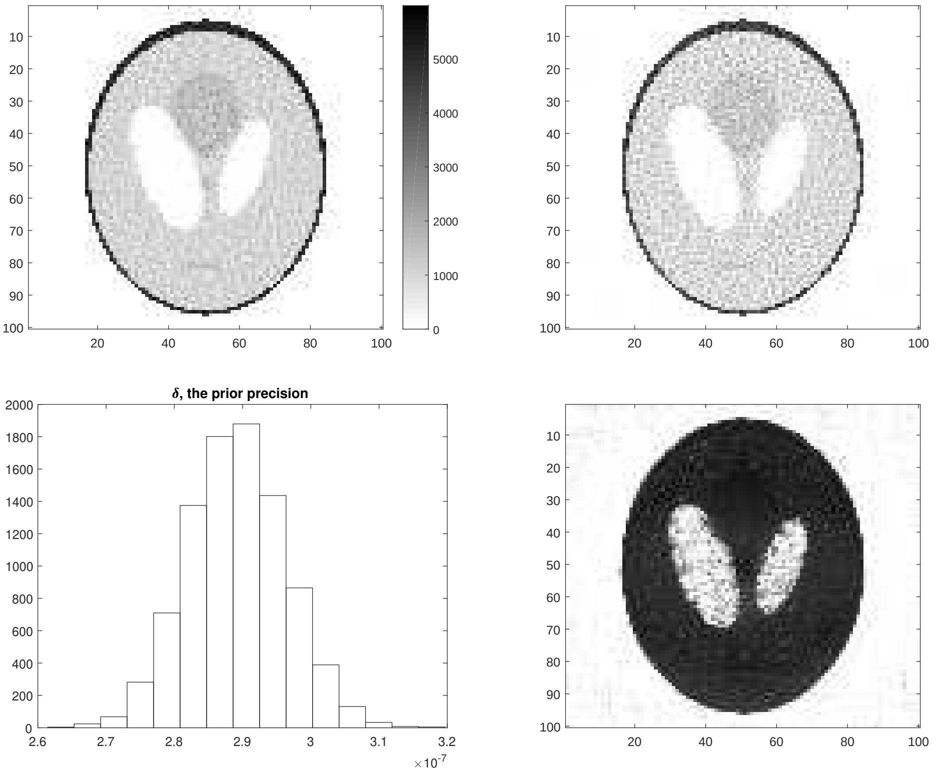

FIG. 10. *The upper-left plot is of the mean of the image samples. The upper-right plot is the MAP estimator with δ taken to be the mean of the δ samples. In the lower left is a histogram of the δ (regularization parameter) samples. In the lower right is the pixelwise standard deviation image.*

implementing these MCMC methods on test cases from image deblurring and computed tomography.

6. Acknowledgments. We would like to acknowledge the referees, whose comments and suggestions helped to improve the work and the presentation.

REFERENCES

[1] S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart, *Analysis of the Gibbs sampler for hierarchical inverse problems*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 511–544.

[2] J. M. Bardsley, *Computational Uncertainty Quantification for Inverse Problems*, SIAM, Philadelphia, 2018.

[3] J. M. Bardsley, *MCMC-based image reconstruction with uncertainty quantification*, SIAM J. Sci. Comput., 34 (2012), pp. A1316–A1332.

[4] J. M. Bardsley and C. Fox, *An MCMC method for uncertainty quantification in nonnegativity constrained inverse problems*, Inverse Probl. Sci. Eng., 20 (2012), pp. 477–498.

[5] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910.

[6] J. M. Bardsley and C. R. Vogel, *A nonnnegatively constrained convex programming method for image reconstruction*, SIAM J. Sci. Comput., 25(4) (2004), pp. 1326–1343.

[7] N. Biswas and P. E. Jacob, *Estimating Convergence of Markov Chains with L-Lag Couplings*, https://arxiv.org/abs/1905.09971.

[8] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, Springer, Berlin, 2008.

[9] D. Calvetti, A. Kuceyeski, and E. Somersalo, *Sampling-based analysis of a spatially distributed model for liver metabolism at steady state*, SIAM J. Multiscale Model. Simulation, 7 (2008), pp. 407–431.

[10] D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing*, Springer, Berlin, 2007.

[11] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.

[12] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., CRC Press, Boca Raton, FL, 2013.

[13] J. Geweke, *Evaluating the accuracy of sampling-based approaches to calculation of posterior moments*, Bayesian Statist., 4 (1992), pp. 169–193.

[14] J. Gorham and L. Mackey, *Measuring sample quality with Stein's method*, in Proceedings 28th International Conference on Neural Information Processing Systems—Volume 1, NIPS'15 (2015), MIT Press, Cambridge, MA, pp. 226–234.

[15] P. Green, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Trans. Med. Imaging, 9 (1990), pp. 84–93.

[16] H. Ishwaran and J. S. Rao, *Spike and slab variable selection: Frequentist and Bayesian strategies*, Ann. Statist., 33 (2005), pp. 730–773.

[17] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, Philadelphia, 2010.

[18] L. R. Lines and R. T. Newrick, *Fundamentals of Geophysical Interpretation*, Society of Exploration Geophysicists, Houston, TX, 2004.

[19] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, Berlin, 2005.

[20] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, *Sampling via measure transport: An introduction*, in Handbook of Uncertainty Quantification, R. Ghanem, D. Higdon, and H. Owhadi, eds., Springer, Cham, Switzerland, 2016, pp. 1–41.

[21] A. M. Michalak and P. K. Kitanidis, *A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification*, Water Resour. Res., 39 (2003), 1033.

[22] A. M. Michalak, *Gibbs sampler for inequality-constrained geostatistical interpolation and inverse modelling*, Water Resour. Res., 44 (2008), W09437.

[23] J. J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM J. Optim., 1 (1991), pp. 93–113.

[24] J. M. Ollinger and J. A. Fessler, *Positron-emission tomography*, IEEE Signal Process. Mag., January 1997, pp. 43–55.

[25] V. Rocková and E. I. George, *The spike-and-slab LASSO*, J. Amer. Statist. Assoc., 113 (2018), pp. 431–444.

[26] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, Berlin, 2004.

[27] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[28] D. L. Snyder, A. M. Hammoud, and R. L. White, *Image recovery from data acquired with a charge-coupled-device camera*, J. Opt. Soc. Amer. A, 10 (1993), pp. 1014–1023.

[29] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.