Density Deconvolution with Normalizing Flows

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Density Deconvolution with Normalizing Flows

Tim Dockhorn * 1 James A. Ritchie * 2 Yaoliang Yu 1 Iain Murray 2

Abstract
Density deconvolution is the task of estimating a probability density function given only noise-corrupted samples. We can fit a Gaussian mixture model to the underlying density by maximum likelihood if the noise is normally distributed, but would like to exploit the superior density estimation performance of normalizing flows and allow for arbitrary noise distributions. Since both adjustments lead to an intractable likelihood, we resort to amortized variational inference. We demonstrate some problems involved in this approach, however, experiments on real data demonstrate that flows can already out-perform Gaussian mixtures for density deconvolution.

1. Introduction
Density estimation is the fundamental statistical task of estimating the density of a distribution given a finite set of measurements. However, in many scientific fields (see examples in Carroll et al., 2006), one only has access to a noise-corrupted set of measurements. Given knowledge of the statistics of the noise, density deconvolution methods attempt to recover the density function of the unobserved noise-free samples rather than the noisy measurements.

In this work we consider the problem of additive noise, where observed samples \( \{w_i\}_{i=1}^m \) are produced by adding independent noise to unobserved values \( \{v_i\}_{i=1}^m \):

\[
w_i = v_i + n_i.
\]  (1)

We also assume that the density function of the noise \( p_n(n_i) \) is perfectly known for every observation. The density function of observations \( w_i \) is then a convolution

\[
p(w_i) = \int_v p_n(v_i - v) p(v) \, dv.
\]  (2)

When the noise distribution is constant, we could estimate the density of the observations \( w \) with any density estimator and then solve Equation (2), e.g. with a kernel density estimator using Fourier transforms (e.g., Liu & Taylor, 1989; Carroll & Hall, 1988; Fan, 1991; Devroye, 1989) or wavelet decompositions (Pensky & Vidakovic, 1999).

When the noise distribution is different for each observation, we only have one sample from each convolved density. This extreme deconvolution setting (Bovy et al., 2011) has previously been tackled by fitting a Gaussian Mixture Model (GMM) to the underlying density \( p(v) \). When the noise distributions are all Gaussian, the marginal likelihood \( p(w_i) \) is tractable, and the GMM can be fitted by Expectation-Maximisation (EM, Bovy et al., 2011) or Stochastic Gradient Descent (SGD, Ritchie & Murray, 2019).

Given enough mixture components, any density function can be approximated arbitrarily closely using GMMs. In practice, however, other representations of densities can be easier to fit, and often generalize better. There is growing interest in normalizing flows (Tabak & Vanden-Eijnden, 2010; Tabak & Turner, 2013), a class of methods that transform a simple source density into a complex target density. Normalizing flows are an efficient alternative to GMMs (e.g., Rezende & Mohamed, 2015), providing both good scalability and high expressivity, and have shown promise in applications similar to density deconvolution (Cranmer et al., 2019).

In this work, we model the underlying density \( p(v) \) with a normalizing flow. The marginal likelihood \( p(w_i) \) is intractable, so we resort to approximate inference. We use amortized variational inference (Section 2), closely following Variational Auto-Encoders (VAEs, Kingma & Welling, 2014; Rezende et al., 2014). Unlike for VAEs, in our framework, the model between the latent \( v \) and observed \( w \) vectors is fixed.

In this proof of concept, we use a fixed Gaussian noise distribution, but our approach would also allow us to use arbitrary and varying noise distributions, as found in realistic applications (e.g., Anderson et al., 2018). In a setting well-suited to the existing Gaussian mixture approach, we find that fitting
Density Deconvolution with Normalizing Flows

flows is harder (Section 4.1), possibly motivating further work on approximate inference in this setting. Nevertheless, on real data, we demonstrate that flows can already outperform GMMs for density deconvolution (Section 4.2).

2. Methods

We take a variational approach (Jordan et al., 1999) to density deconvolution. Introducing an approximate posterior $q_\phi(v)$ gives a lower bound to the log-marginal likelihood

$$\log p(w_i) = \log \int_v p_n(w_i - v) p_\theta(v) \, dv$$

(3)

$$= \log \int_v p_n(w_i - v) \frac{q_\phi(v)}{q_\phi(v)} \, dv$$

(4)

$$\geq E_q[\log p_n(w_i - v)] - D_{KL}(q_\phi(v) \parallel p_\theta(v)) = L,$$

(5)

where $L$ is the evidence lower bound (ELBO, see Appendix A). Our approximate posterior $q_\phi(v)$, a recognition network, represents beliefs about an underlying value $v$ given an observation $w$ and the parameters of the noise.

The ELBO gives a unified objective for both the parameters $\theta$ of the model and the parameters $\phi$ of the recognition network. Stochastic gradient descent only needs unbiased estimates of the ELBO, which we obtain by Monte Carlo

$$L \approx L(K) = \frac{1}{K} \sum_{k=1}^{K} \log \left[ \frac{p_n(w_i - v_k) p_\theta(v_k)}{q_\phi(v_k)} \right],$$

(6)

where $K$ Monte Carlo samples are simulated $v_k \sim q_\phi(v)$.

Our variational approach follows that of Variational Auto-Encoders (VAEs, Kingma & Welling, 2014; Rezende et al., 2014), which provide a framework for amortized variational inference in graphical models. The focus of VAEs, however, is usually to build generative models matching the observations $w$. In contrast, our main target is estimating an underlying density function $p(v)$. For certain applications, e.g. denoising a noisy measurement, we may also be interested in the approximate posterior $q_\phi(v)$. Another difference between our method and VAEs is that we do not learn a likelihood model between the latent variables $v$ and the observations $w$. Instead, our “likelihood model” is fully-characterized by the problem itself as $p_n(w_i - v)$.

We model both $p_\theta(v)$ and $q_\phi(v)$ as normalizing flows. Normalizing flows model probability density functions by transforming a source density $\pi(u)$ into a target density $\hat{\pi}(v)$ using an invertible, differentiable transformation $T$

$$v = T(u).$$

(7)

The density of $v$ can be computed using the change-of-variable formula

$$\hat{\pi}(v) = \pi(T^{-1}(v)) \left| \frac{\partial}{\partial v} (T^{-1}(v)) \right|$$

(8)

In particular, we model $p_\theta(v)$ and $q_\phi(v)$ with autoregressive flows. For $p_\theta(v)$ we use a Masked Autoregressive Flow (MAF, Papamakarios et al., 2017), where a single neural network pass can compute $T^{-1}(v)$, and therefore the densities $\hat{\pi}(v)$ required during training. Inverting the network, to generate samples, requires $D$ neural network passes for $D$-dimensional data. For $q_\phi(v)$ we use the same network architecture to represent $T(u)$, corresponding to an Inverse Autoregressive Flow (IAF, Kingma et al., 2016). During training, this choice gives fast one-pass generation of samples with their densities. For an extensive review on normalizing flows, we refer the reader to Papamakarios et al. (2019).

3. Related Work

Importance weighting: The Importance Weighted Autoencoder (Burda et al., 2015) has the same architecture as the standard VAE, however, it is trained on a lower bound that is tighter than the standard ELBO. Applying this idea to our model results in the following lower bound:

$$\log p(w_i) \geq \log \left[ \frac{1}{K} \sum_{k=1}^{K} \frac{p_n(w_i - v_k) p_\theta(v_k)}{q_\phi(v_k)} \right]$$

(9)

$$= L_{IW}(K).$$

(10)

It can be shown (Cremer et al., 2017) that, in expectation, $L_{IW}(K)$ is equivalent to $L(K)$ with an implicit, more expressive approximate posterior. A theoretical advantage of $L_{IW}(K)$ over $L(K)$ is that the former is consistent (under some mild boundedness assumptions, Burda et al., 2015, Theorem 1), i.e., $\lim_{K \to \infty} L_{IW}(K) = \log p(w_i)$.

In fact, it is possible to construct an unbiased estimator of $\log p(w_i)$ using $L_{IW}(K)$ with finite $K$ (Luo et al., 2020), in combination with a Russian Roulette Estimator (Kahn, 1955). A drawback of this approach is that there is no guarantee that the variance of the estimator is finite.

Inference suboptimality: When using the ELBO, or $L_{IW}(K)$ with finite $K$, we only have a bound on the marginal log-likelihood $\log p(w_i)$. This bound is loose when the approximate posterior is incorrect, which happens either because the form of the posterior cannot be represented, or because the recognition network does not produce good variational parameters for all data points. Both issues can be overcome by choosing the approximate posterior from an expressive variational family (Cremer et al., 2018), which is why we use a flow.

Expressive priors for representation learning: The standard VAE has a fixed prior, usually a multivariate standard normal distribution, but VAEs with more expressive priors have been proposed. Expressive priors are particularly useful when the distribution of the latent variables is used for representation learning.
A simple generalization for the prior is a learnable GMM (e.g. Nalisnick et al., 2016; Dilokthanakul et al., 2016), which in our context would result in the existing extreme deconvolution model (Bovy et al., 2011), with no need for variational inference. Another approach is to model the prior with a collection of categorical distributions (e.g. van den Oord et al., 2017), which would be appropriate if an observation is well-modeled as a composition of prototype sources. We use MAF, because for the applications we have in mind (e.g., Anderson et al., 2018), we want to use a flexible, continuous prior representation. VAEs have also used autoregressive flow priors before (Chen et al., 2017).

4. Experiments

In this section, we compare our method to a baseline of GMMs fitted with the Extreme Deconvolution (XD) model (Bovy et al., 2011), as by Ritchie & Murray (2019).

4.1. Mixture of Gaussians

In this synthetic task, the target underlying density is a mixture of Gaussians. We fit the models from observations that include additional noise from a Gaussian with fixed covariance. Figure 1 shows density plots of both the latent data v and the observed data w.

The exact posterior for a latent datapoint given a noisy observation is itself a mixture of Gaussians, and for some observations the components may be highly isolated. We have deliberately picked an example where the prior \( p(v) \) should be reasonably easy for a flow to model, but the posterior \( p(v | w) \) may cause issues for flows, as they are known to have trouble fitting mixture of Gaussians when the components are well-separated (e.g., Jaini et al., 2019). Full experimental details are reported in Appendix B.1

Table 1 reports test average negative log-likelihood on both v and w, referred to as \( -\log p(v) \) and \( -\log p(w) \), respectively. We estimate \( \log p(w) \) using \( L_{IW}(100) \). The GMM, the true model class, has the best results for both v and w. The flows give quite close estimates for \( \log p(w) \) when \( K > 1 \) for both \( L \) and \( L_{IW} \), but show very high variance in their estimates of \( \log p(v) \) relative to the variance for \( \log p(w) \).

The top row of Figure 2 shows example density plots using samples from the priors of our fitted models. The fitted GMM has matched the ground truth GMM closely. The flows recover the broad shape of the ground truth model, but those trained with \( L(1) \) and \( L(50) \) put too much mass in the centre. The flow trained with \( L_{IW}(50) \) matched the Gaussian mixture model on the run shown, but the results had high variance, and other runs do not look as good.

The bottom row of Figure 2 shows example posteriors for the fitted models. The exact posterior for the GMM has two isolated modes, and is a close match to the ground truth posterior. The approximate posteriors for the flows trained with \( L(1) \) and \( L(50) \) are not good matches to the GMM posterior, as neither have isolated modes, but are reasonably consistent with their corresponding priors. The approximate posterior for the flow trained with \( L_{IW}(50) \) uses samples drawn from \( q_{\theta} \) with sampling-importance-resampling (e.g., Rubin, 1988). The resampling reflects this recognition network’s role as an adaptive proposal distribution under the \( L_{IW} \) objective rather than a direct approximation to the posterior (Cremer et al., 2017). This reweighted approximation is a much better match to the ground truth posterior, but as with the prior, the variance across training runs was high, and other examples are qualitatively worse.

To establish whether the inability of our procedure to consistently recover the correct ground truth model is a problem with the prior flow itself, the approximate posterior, the interaction of both, or the training objectives, we tried various methods of training each part separately. All results for these experiments are summarized in Table 2. Additional density plots are also available in Appendix C.

The flow \( p_{\theta}(v) \) was pretrained directly on the noise free samples underlying the training set via maximum likelihood. The test average negative log-likelihood on v is much closer to the GMM. Therefore, while model mismatch is a slight
disadvantage for the flows here, it is not entirely responsible for their worse results. Similarly, we trained the recognition network directly on noisy samples from the training set, paired with samples from the exact ground truth posterior. When combined with the pretrained prior, the test likelihood on \( w \) was much closer to that of the GMM, suggesting that the flows can represent useful posteriors.

We then fitted the models with the \( \mathcal{L}(50) \) objective, but initialized with the pretrained prior and posterior. The variational objective was significantly improved by moving to a model with similar \( \log p(w) \) as the GMM, however, doing so made \( \log p(v) \) worse. Despite using fairly flexible flows, the variational bound is not tight, and biases us towards worse prior models.

Finally we tried fitting a GMM with the \( \mathcal{L}(50) \) objective rather than by maximizing the log-likelihood directly, using both samples from the exact posterior and samples from the conditional flow approximate posterior. When using exact samples, the variational bound is tight, but we experience the noisier gradients of variational fitting: the GMM still recovers the same result as before. However, using posterior samples from the flow causes a similar bias to before, showing that the flows are not learning good enough posteriors for variational inference to be accurate.

### 4.2. UCI datasets

We now compare the two methods on two small UCI datasets (Dua & Graff, 2017) that are difficult to fit with GMMs (Uria et al., 2013). We discarded discrete-valued attributes and normalized the data. The datasets are then split into training and testing sets; 90% are used for training and 10% are used for testing. We subsequently add noise from a zero-mean independent normal distribution with diagonal covariance matrix \( \Sigma_{ii} = 0.1 \) to each noise-free training point to generate the observations \( w_i \).

For our method, we use the objective \( \mathcal{L}_{IW}(50) \) as it yielded the best test average negative log-likelihood in the previous section. We note, however, that this might not be the best choice as the high-variance pattern for \( \log p(v) \) (see Table 1) might persist. The test results are reported in Table 3; full experimental details can be found in Appendix B.2. Flows outperformed GMMs in both cases.

As an ablation, we tried fitting conventional flows to the noisy observations \( w_i \), without correcting for the noise in any way (Table 4). These flows beat the GMMs on both datasets, showing the importance of using good representations, however, the results are still significantly worse than flows with deconvolution (right column Table 3).
5. Conclusion

In this preliminary work, we have outlined an approach for density deconvolution using normalizing flows and arbitrary noise distributions by turning the deconvolution problem into an approximate inference problem. Our experiments on a toy setup have shown that variational inference with an inaccurate posterior can prevent the model prior from learning the underlying noise-free density. For future work, we are planning to experiment with unbiased inference, e.g., using Markov chain Monte Carlo methods (e.g., Glynn & Rhee, 2014; Qiu et al., 2020) or importance weighting in combination with the Russian Roulette Estimator (Luo et al., 2020). Nevertheless, we have already been able to demonstrate that normalizing flows fitted with our approach can beat GMMs for density deconvolution on (small) real-world datasets, which indicates that further research on how to fit normalizing flows in this context is worth pursuing.

Acknowledgements

We are grateful to Artur Bekasov and Conor Durkan for discussions around choices of normalizing flow for this application and use of their flows library (Durkan et al., 2019). Our experiments also made use of corner.py (Foreman-Mackey, 2016), Matplotlib (Hunter, 2007), NumPy (Oliphant, 2006), Pandas (McKinney, 2010) and PyTorch (Paszke et al., 2019). This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We gratefully acknowledge funding support from NSERC and the Canada CIFAR AI Chairs Program.

References


Density Deconvolution with Normalizing Flows


### A. The Evidence Lower Bound

The KL divergence between the variational distribution and the posterior can be written as

\[
D_{\text{KL}}(q_\phi(v) \parallel p(v | w_i)) = \mathbb{E}_q \left[ \log q_\phi(v) - \log p_\theta(v | w_i) \right] \tag{11}
\]

\[
= \mathbb{E}_q \left[ \log q_\phi(v) - \log p_\theta(v, w_i) \right] + \log p_\theta(w_i). \tag{12}
\]

The joint distribution of \(v\) and \(w\) can be computed as

\[
p_\theta(v, w_i) = \int p_\theta(v, w_i, n_i) \, dn_i \tag{13}
\]

\[
= \int n_i \, \delta(w_i = v + n_i) p_\theta(v, n_i) \, dn_i \tag{14}
\]

\[
= p_\theta(v) p_{n_i}(w_i - v), \tag{15}
\]

where \(\delta(\cdot)\) is the Dirac delta distribution. Hence,

\[
\log p_\theta(w_i) = D_{\text{KL}}(q_\phi(v) \parallel p_\theta(v | w_i)) + \mathbb{E}_q \left[ \log p_{n_i}(w_i - v) \right] - D_{\text{KL}}(q_\phi(v) \parallel p_\theta(v)) \tag{16}
\]

\[
= D_{\text{KL}}(q_\phi(v) \parallel p_\theta(v | w_i)) + \mathcal{L}. \tag{17}
\]

### B. Details for Experiments

All code used to run the experiments is available: [https://github.com/bayesiains/density-deconvolution](https://github.com/bayesiains/density-deconvolution)

#### B.1. Mixture of Three Gaussians

Latent datapoints \(v_i\) were drawn from a mixture of 3 Gaussians, with equal mixture weights, means

\[
m_1 = [-2 \ 0]^T, \quad m_2 = [0 \ -2]^T, \quad m_3 = [0 \ 2]^T, \tag{18}
\]

and covariances

\[
C_1 = \begin{bmatrix} 0.3^2 & 0 \\ 0 & 1 \end{bmatrix}, \quad C_2 = C_3 = \begin{bmatrix} 1 & 0 \\ 0 & 0.3^2 \end{bmatrix}. \tag{19}
\]

Zero-mean Gaussian noise with covariance

\[
S = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{20}
\]

was added to each \(v_i\) to produce \(w_i\). Training and test sets consisted of 50 000 samples each, whilst the validation set had 12 500 samples.

### B.2. UCI datasets

Since the datasets are relatively small, we tune the hyperparameters of the models using 5-fold cross-validation and grid search; the parameters of the grid search are reported in Table 5. Once the hyperparameter values had been determined, we trained the models using a tenth of the training data for early-stopping and measured their performance on the 10% held-out test data.

In contrast to the setup in Appendix B.1, we used simple dense layers rather than residual layers within each MAF.
Density Deconvolution with Normalizing Flows

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Tested values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed learning rate</td>
<td>0.001*, 0.0005, 0.0001</td>
</tr>
<tr>
<td>MAF layers (p_\theta(v))</td>
<td>3*, 4, 5†</td>
</tr>
<tr>
<td>MAF layers (q_\phi(v))</td>
<td>3, 4*, 5†</td>
</tr>
<tr>
<td>MAF hidden features</td>
<td>64, 128*, 5†</td>
</tr>
<tr>
<td>MAF hidden blocks</td>
<td>1*, 2†</td>
</tr>
<tr>
<td>Fixed learning rate</td>
<td>0.01*, 0.005†, 0.001</td>
</tr>
<tr>
<td># of mixture components</td>
<td>20†., 50, 100, 200, 300*</td>
</tr>
</tbody>
</table>

Table 5. Tested values for hyperparameter tuning on UCI datasets. The chosen values for each dataset are marked as: (*) White wine, (†) Red wine.

layer. Masking of the dense layers was done using the standard MADE architecture (Germain et al., 2015). A further difference is that we conditioned the recognition network \(q_\phi(v)\) only on \(w\).

Training ran until no improvement in validation loss was observed for 30 epochs. For this experiment, we did not apply any Dropout. The minibatch size was chosen to be 100.

C. Additional Plots

Figure 3. Density plots of pretrained models. Left: Flow trained directly on noise-free samples. Right: Posterior flow trained as a conditional density estimator on pairs of noisy observations and samples from the exact posterior. The trail linking the posterior modes does not have a large penalty under our objective function.

Figure 4. Density plots of models initialized with pretrained flows, then trained jointly with \(\mathcal{L}(50)\). Left: Prior \(p_\theta(v)\). Right: Posterior \(q_\phi(v)\) for a given test point. The trail between the posterior modes has been reduced, but is still present. The fitted prior density has gotten slightly worse (Table 2).

Figure 5. Density plots of a GMM fitted with the \(\mathcal{L}(50)\) objective using a conditional-flow posterior. Left: Prior \(p_\theta(v)\). Right: Posterior \(q_\phi(v)\) for a given test point. We are using the ground-truth model class, but the fit of the prior is not as good as when using the exact posterior. Importance weighted training may help remove the mass between the modes (Figure 2, right), but we did not get stable results (Table 1).

Figure 6. Density plots of a GMM fitted with the \(\mathcal{L}(50)\) objective using the exact posterior. Left: Prior \(p_\theta(v)\). Right: Posterior \(q_\phi(v)\) for a given test point.