ON THE POSTERIOR DISTRIBUTION IN DENOISING: APPLICATION TO UNCERTAINTY QUANTIFICATION SUPPLEMENTARY MATERIAL

CONTENTS

A Proof of Theorem 1

We start with the case $k \geq 2$ (bottom two lines in (4)). In this case, the conditional moment $\mu_k(y)$ can be expressed using Bayes' formula as

$$\begin{aligned}
\mu_k(y) &= \mathbb{E}\left[(\mathrm{x}-\mu_1(\mathrm{y}))^k \,\middle|\, \mathrm{y}=y\right] \\
&= \int (x-\mu_1(y))^k p_{\mathrm{x}|\mathrm{y}}(x|y)dx \\
&= \frac{\int (x-\mu_1(y))^k p_{\mathrm{y}|\mathrm{x}}(y|x)p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)} \\
&= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int (x-\mu_1(y))^k \exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)}. \qquad \text{(S1)}
\end{aligned}$$

Denoting the numerator by $q(y) \triangleq (2\pi\sigma^2)^{-\frac{1}{2}}\int (x-\mu_1(y))^k \exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx$, we can write the derivative of $\mu_k(y)$ as

$$\begin{aligned}
\mu_k'(y) &= \frac{q'(y)p_{\mathrm{y}}(y) - q(y)p_{\mathrm{y}}'(y)}{p_{\mathrm{y}}^2(y)} \\
&= \frac{q'(y)}{p_{\mathrm{y}}(y)} - \frac{q(y)}{p_{\mathrm{y}}(y)}\frac{p_{\mathrm{y}}'(y)}{p_{\mathrm{y}}(y)} \\
&= \frac{q'(y)}{p_{\mathrm{y}}(y)} - \mu_k(y)\frac{p_{\mathrm{y}}'(y)}{p_{\mathrm{y}}(y)} \\
&= \frac{q'(y)}{p_{\mathrm{y}}(y)} - \mu_k(y)\frac{d\log p_{\mathrm{y}}(y)}{dy} \\
&= \frac{q'(y)}{p_{\mathrm{y}}(y)} - \frac{1}{\sigma^2}\mu_k(y)(\mu_1(y)-y), \qquad \text{(S2)}
\end{aligned}$$

where we used the fact that $\frac{d\log p_{\mathrm{y}}(y)}{dy} = \frac{1}{\sigma^2}(\mu_1(y)-y)$ (see *e.g.*, (Efron, 2011; Miyasawa et al., 1961; Stein, 1981)). The first term in this expression is given by

$$\begin{aligned}
\frac{q'(y)}{p_{\mathrm{y}}(y)} &= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int \frac{d}{dy}\left[(x-\mu_1(y))^k \exp\{-\frac{1}{2\sigma^2}(y-x)^2\}\right]p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)} \\
&= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int \left(-k(x-\mu_1(y))^{k-1}\mu_1'(y) - (x-\mu_1(y))^k\frac{1}{\sigma^2}(y-x)\right)\exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)} \\
&= \frac{\int \left(-k(x-\mu_1(y))^{k-1}\mu_1'(y) - (x-\mu_1(y))^k\frac{1}{\sigma^2}(y-x)\right)p_{\mathrm{y}|\mathrm{x}}(y|x)p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)} \\
&= \int \left(-k(x-\mu_1(y))^{k-1}\mu_1'(y) - (x-\mu_1(y))^k\frac{1}{\sigma^2}(y-x)\right)p_{\mathrm{x}|\mathrm{y}}(x|y)dx \\
&= \mathbb{E}\left[-k(\mathrm{x}-\mu_1(\mathrm{y}))^{k-1}\mu_1'(\mathrm{y}) - (\mathrm{x}-\mu_1(\mathrm{y}))^k\frac{1}{\sigma^2}(\mathrm{y}-\mathrm{x}) \,\middle|\, \mathrm{y}=y\right]. \qquad \text{(S3)}
\end{aligned}$$

To allow unified treatment of the cases $k=2$ and $k>2$, let us denote

$$\psi_k(y) \triangleq \mathbb{E}\left[(\mathrm{x}-\mu_1(\mathrm{y}))^k \,\middle|\, \mathrm{y}=y\right] = \begin{cases} 0 & k=1, \\ \mu_k(y) & k \geq 2. \end{cases} \qquad \text{(S4)}$$

We therefore have

$$\begin{aligned}
\frac{q'(y)}{p_{\mathrm{y}}(y)} &= -k\psi_{k-1}(y)\mu_1'(y) - \frac{1}{\sigma^2}\psi_k(y)y + \frac{1}{\sigma^2}\mathbb{E}\left[(\mathrm{x}-\mu_1(\mathrm{y}))^k\mathrm{x} \,\middle|\, \mathrm{y}=y\right] \\
&= -k\psi_{k-1}(y)\mu_1'(y) - \frac{1}{\sigma^2}\psi_k(y)y + \frac{1}{\sigma^2}\mathbb{E}[(\mathrm{x}-\mu_1(\mathrm{y}))^k(\mathrm{x}-\mu_1(\mathrm{y})+\mu_1(\mathrm{y})) \,|\, \mathrm{y}=y] \\
&= -k\psi_{k-1}(y)\mu_1'(y) - \frac{1}{\sigma^2}\psi_k(y)y + \frac{1}{\sigma^2}\left(\psi_{k+1}(y)+\psi_k(y)\mu_1(y)\right) \\
&= -k\psi_{k-1}(y)\mu_1'(y) + \frac{1}{\sigma^2}\psi_{k+1}(y) + \frac{1}{\sigma^2}\psi_k(y)\left(\mu_1(y)-y\right). \qquad \text{(S5)}
\end{aligned}$$

Substituting this back into (S2), we obtain that

$$\begin{aligned}\mu_k'(y) &= -k\psi_{k-1}(y)\mu_1'(y) + \frac{1}{\sigma^2}\psi_{k+1}(y) + \frac{1}{\sigma^2}\psi_k(y)\left(\mu_1(y)-y\right) - \frac{1}{\sigma^2}\mu_k(y)\left(\mu_1(y)-y\right)\\ &= -k\psi_{k-1}(y)\mu_1'(y) + \frac{1}{\sigma^2}\psi_{k+1}(y),\end{aligned} \tag{S6}$$

where we used the fact that $\psi_k(y) = \mu_k(y)$ for all $k \geq 2$. Now, for $k = 2$ this equation reads

$$\mu_2'(y) = \frac{1}{\sigma^2}\mu_3(y), \tag{S7}$$

and for $k \geq 3$, it reads

$$\mu_k'(y) = -k\mu_{k-1}(y)\mu_1'(y) + \frac{1}{\sigma^2}\mu_{k+1}(y). \tag{S8}$$

We thus have that

$$\begin{aligned}\mu_3(y) &= \sigma^2\mu_2'(y),\\ \mu_{k+1}(y) &= \sigma^2\mu_k'(y) + k\sigma^2\mu_{k-1}(y)\mu_1'(y), \quad k \geq 3.\end{aligned} \tag{S9}$$

Note that an equivalent expression for the last line is obtained by replacing $\sigma^2\mu_1'(y)$ with $\mu_2(y)$, as we prove below. This completes the proof for $k \geq 2$.

The case $k = 1$ can be treated similarly. Here,

$$\begin{aligned}\mu_1(y) &= \mathbb{E}[\mathrm{x} \,|\, \mathrm{y} = y]\\ &= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int x\exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)},\end{aligned} \tag{S10}$$

so that we define $q(y) \triangleq (2\pi\sigma^2)^{-\frac{1}{2}}\int x\exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx$. We thus have

$$\begin{aligned}\frac{q'(y)}{p_{\mathrm{y}}(y)} &= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int \frac{d}{dy}\left[x\exp\{-\frac{1}{2\sigma^2}(y-x)^2\}\right]p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)}\\ &= \frac{(2\pi\sigma^2)^{-\frac{1}{2}}\int \frac{1}{\sigma^2}(x-y)\exp\{-\frac{1}{2\sigma^2}(y-x)^2\}p_{\mathrm{x}}(x)dx}{p_{\mathrm{y}}(y)}\\ &= \frac{1}{\sigma^2}\mathbb{E}[\mathrm{x}(\mathrm{x}-\mathrm{y}) \,|\, \mathrm{y} = y]\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathrm{x}^2 \,|\, \mathrm{y} = y] - \mu_1(y)y\right).\end{aligned} \tag{S11}$$

Therefore,

$$\begin{aligned}\mu_1'(y) &= \frac{q'(y)}{p_{\mathrm{y}}(y)} - \frac{1}{\sigma^2}\mu_1(y)(\mu_1(y)-y)\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathrm{x}^2 \,|\, \mathrm{y} = y] - \mu_1(y)y\right) - \frac{1}{\sigma^2}\mu_1(y)(\mu_1(y)-y)\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathrm{x}^2 \,|\, \mathrm{y} = y] - \mu_1^2(y)\right)\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathrm{x}^2 \,|\, \mathrm{y} = y] - \mathbb{E}[\mathrm{x} \,|\, \mathrm{y} = y]^2\right)\\ &= \frac{1}{\sigma^2}\mu_2(y),\end{aligned} \tag{S12}$$

which demonstrates that

$$\mu_2(y) = \sigma^2\mu_1'(y). \tag{S13}$$

This completes the proof for $k = 1$.

B Proof of Theorem 2

We begin with the case $k = 1$ (first line in (10)), by directly deriving the matrix form (11). Using Bayes' formula, the posterior mean $\boldsymbol{\mu}_1(\boldsymbol{y})$ can be expressed as

$$\begin{aligned}\boldsymbol{\mu}_1(\boldsymbol{y}) &= \mathbb{E}[\mathrm{x}|\mathrm{y}=\boldsymbol{y}]\\ &= \int_{\mathbb{R}^d} \boldsymbol{x} p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}\\ &= \frac{\int_{\mathbb{R}^d} \boldsymbol{x} p_{\mathrm{y}|\mathrm{x}}(\boldsymbol{y}|\boldsymbol{x}) p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}}{p_{\mathrm{y}}(\boldsymbol{y})}\\ &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \boldsymbol{x} \exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\} p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}}{p_{\mathrm{y}}(\boldsymbol{y})}.\end{aligned} \tag{S14}$$

Therefore, denoting the numerator by $q(\boldsymbol{y}) \triangleq \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \boldsymbol{x} \exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\} p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}$, we can write the Jacobian of $\boldsymbol{\mu}_1$ at $\boldsymbol{y}$ as

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}(\boldsymbol{y})}{\partial \boldsymbol{y}} &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}} p_{\mathrm{y}}(\boldsymbol{y}) - q(\boldsymbol{y}) \left(\nabla p_{\mathrm{y}}(\boldsymbol{y})\right)^\top}{p_{\mathrm{y}}^2(\boldsymbol{y})}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}}}{p_{\mathrm{y}}(\boldsymbol{y})} - \frac{q(\boldsymbol{y})}{p_{\mathrm{y}}(\boldsymbol{y})} \frac{\left(\nabla p_{\mathrm{y}}(\boldsymbol{y})\right)^\top}{p_{\mathrm{y}}(\boldsymbol{y})}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}}}{p_{\mathrm{y}}(\boldsymbol{y})} - \boldsymbol{\mu}_1(\boldsymbol{y}) \frac{\left(\nabla p_{\mathrm{y}}(\boldsymbol{y})\right)^\top}{p_{\mathrm{y}}(\boldsymbol{y})}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}}}{p_{\mathrm{y}}(\boldsymbol{y})} - \boldsymbol{\mu}_1(\boldsymbol{y}) \left(\nabla \log p_{\mathrm{y}}(\boldsymbol{y})\right)^\top\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}}}{p_{\mathrm{y}}(\boldsymbol{y})} - \frac{1}{\sigma^2} \boldsymbol{\mu}_1(\boldsymbol{y}) \left(\boldsymbol{\mu}_1(\boldsymbol{y})^\top - \boldsymbol{y}^\top\right).\end{aligned} \tag{S15}$$

Here, $\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}} \in \mathbb{R}^{d\times d}$ denotes the Jacobian of $q : \mathbb{R}^d \to \mathbb{R}^d$ at $\boldsymbol{y}$, and we used the fact that $\nabla \log p_{\mathrm{y}}(\boldsymbol{y}) = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1(\boldsymbol{y}) - \boldsymbol{y})$ (Efron, 2011; Miyasawa et al., 1961; Stein, 1981). The first term in (S15) can be further simplified as

$$\begin{aligned}\frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}}}{p_{\mathrm{y}}(\boldsymbol{y})} &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \boldsymbol{x} \exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\} \frac{1}{\sigma^2}(\boldsymbol{x}-\boldsymbol{y})^\top p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}}{p_{\mathrm{y}}(\boldsymbol{y})}\\ &= \frac{\int_{\mathbb{R}^d} \frac{1}{\sigma^2} \boldsymbol{x}(\boldsymbol{x}-\boldsymbol{y})^\top p_{\mathrm{y}|\mathrm{x}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathrm{x}}(\boldsymbol{x}) d\boldsymbol{x}}{p_{\mathrm{y}}(\boldsymbol{y})}\\ &= \int_{\mathbb{R}^d} \frac{1}{\sigma^2} \boldsymbol{x}(\boldsymbol{x}-\boldsymbol{y})^\top p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}\\ &= \frac{1}{\sigma^2} \left(\mathbb{E}[\mathrm{x}\mathrm{x}^\top|\mathrm{y}=\boldsymbol{y}] - \mathbb{E}[\mathrm{x}|\mathrm{y}=\boldsymbol{y}]\, \boldsymbol{y}^\top\right)\\ &= \frac{1}{\sigma^2} \left(\mathbb{E}[\mathrm{x}\mathrm{x}^\top|\mathrm{y}=\boldsymbol{y}] - \boldsymbol{\mu}_1(\boldsymbol{y})\, \boldsymbol{y}^\top\right).\end{aligned} \tag{S16}$$

Substituting (S16) back into (S15), we obtain

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}_1(\boldsymbol{y})}{\partial \boldsymbol{y}} &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top|\mathbf{y}=\boldsymbol{y}] - \boldsymbol{\mu}_1(\boldsymbol{y})\,\boldsymbol{y}^\top\right) - \frac{1}{\sigma^2}\boldsymbol{\mu}_1(\boldsymbol{y})\left(\boldsymbol{\mu}_1(\boldsymbol{y})^\top - \boldsymbol{y}^\top\right)\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top|\mathbf{y}=\boldsymbol{y}] - \boldsymbol{\mu}_1(\boldsymbol{y})\boldsymbol{\mu}_1(\boldsymbol{y})^\top\right)\\ &= \frac{1}{\sigma^2}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top|\mathbf{y}=\boldsymbol{y}] - \mathbb{E}[\mathbf{x}|\mathbf{y}=\boldsymbol{y}]\mathbb{E}[\mathbf{x}|\mathbf{y}=\boldsymbol{y}]^\top\right)\\ &= \frac{1}{\sigma^2}\mathrm{Cov}(\mathbf{x}|\mathbf{y}=\boldsymbol{y})\\ &= \frac{1}{\sigma^2}\boldsymbol{\mu}_2(\boldsymbol{y}).\end{aligned} \tag{S17}$$

This completes the proof for $k=1$.

We now move on to the cases $k=2$ and $k\geq 3$ (second and third lines in (10)). Element $(i_1,\ldots,i_k)$ of the posterior kth order central moment can be expressed as

$$\begin{aligned}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k} &= \mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})\cdots(\mathbf{x}_{i_k} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_k})\;\middle|\;\mathbf{y}=\boldsymbol{y}\right]\\ &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\boldsymbol{x}_{i_1} - [\mu_1(\boldsymbol{y})]_{i_1})\cdots(\boldsymbol{x}_{i_k} - [\mu_1(\boldsymbol{y})]_{i_k})\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}}{p_{\mathbf{y}}(\boldsymbol{y})}\\ &= \frac{q(\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})},\end{aligned} \tag{S18}$$

where $q(\boldsymbol{y}) \triangleq \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\boldsymbol{x}_{i_1} - [\mu_1(\boldsymbol{y})]_{i_1})\cdots(\boldsymbol{x}_{i_k} - [\mu_1(\boldsymbol{y})]_{i_k})\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}$. Therefore, for any $i_{k+1}\in\{1,\ldots,d\}$, the derivative of $[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}$ with respect to $\boldsymbol{y}_{i_{k+1}}$ is given by

$$\begin{aligned}\frac{\partial[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}p_{\mathbf{y}}(\boldsymbol{y}) - q(\boldsymbol{y})\frac{\partial p_{\mathbf{y}}(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}^2(\boldsymbol{y})}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})} - \frac{q(\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})}\frac{\frac{\partial p_{\mathbf{y}}(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})} - [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\frac{\partial \log p_{\mathbf{y}}(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}\\ &= \frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})} - \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\left([\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_{k+1}} - \boldsymbol{y}_{i_{k+1}}\right),\end{aligned} \tag{S19}$$

where in the last line we used the fact that $\nabla\log p_{\mathbf{y}}(\boldsymbol{y}) = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1(\boldsymbol{y}) - \boldsymbol{y})$ (Efron, 2011; Miyasawa et al., 1961; Stein, 1981). The first term here can be written as

$$\begin{aligned}\frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})} &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\int\frac{\partial}{\partial \boldsymbol{y}_{i_{k+1}}}\left[(\boldsymbol{x}_{i_1} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1})\cdots(\boldsymbol{x}_{i_k} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k})\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}\right]p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}}{p_{\mathbf{y}}(\boldsymbol{y})}\\ &= \frac{\int -\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_{k+1}}}(\boldsymbol{x}_{i_2} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_2})\cdots(\boldsymbol{x}_{i_k} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k})\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}}{(2\pi\sigma^2)^{\frac{d}{2}}p_{\mathbf{y}}(\boldsymbol{y})} + \cdots\\ &\quad + \frac{\int -(\boldsymbol{x}_{i_1} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1})\cdots\left(\boldsymbol{x}_{i_{k-1}} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_{k-1}}\right)\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k}}{\partial \boldsymbol{y}_{i_{k+1}}}\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}}{(2\pi\sigma^2)^{\frac{d}{2}}p_{\mathbf{y}}(\boldsymbol{y})}\\ &\quad + \frac{\int(\boldsymbol{x}_{i_1} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1})\cdots(\boldsymbol{x}_{i_k} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k})\frac{1}{\sigma^2}\left(\boldsymbol{x}_{i_{k+1}} - \boldsymbol{y}_{i_{k+1}}\right)\exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{x}\|^2\}p_{\mathbf{x}}(\boldsymbol{x})d\boldsymbol{x}}{(2\pi\sigma^2)^{\frac{d}{2}}p_{\mathbf{y}}(\boldsymbol{y})}.\end{aligned} \tag{S20}$$

Let us treat the cases $k = 2$ and $k \geq 3$ separately. When $k = 2$, the above expression contains precisely three terms, but the first two vanish. Indeed, the first term reduces to $-\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_3}}\mathbb{E}[\mathbf{x}_{i_2} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_2}|\mathbf{y} = \boldsymbol{y}] = -\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_3}}([\boldsymbol{\mu}_1(\mathbf{y})]_{i_2} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_2}) = 0$ and the second term to $-\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_2}}{\partial \boldsymbol{y}_{i_3}}\mathbb{E}[\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1}|\mathbf{y} = \boldsymbol{y}] = -\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_2}}([\boldsymbol{\mu}_1(\mathbf{y})]_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1}) = 0$. Therefore, when $k = 2$ we are left only with the last term, which simplifies to

$$\begin{aligned}
\frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_3}}}{p_{\mathbf{y}}(\boldsymbol{y})} &= \frac{1}{\sigma^2}\mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})(\mathbf{x}_{i_2} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_2})(\mathbf{x}_{i_3} - \mathbf{y}_{i_3}) \,\middle|\, \mathbf{y} = \boldsymbol{y}\right] \\
&= \frac{1}{\sigma^2}\mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})(\mathbf{x}_{i_2} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_2})\mathbf{x}_{i_3} \,\middle|\, \mathbf{y} = \boldsymbol{y}\right] - \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}\boldsymbol{y}_{i_3} \\
&= \frac{1}{\sigma^2}\mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})(\mathbf{x}_{i_2} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_2})(\mathbf{x}_{i_3} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_3} + [\boldsymbol{\mu}_1(\mathbf{y})]_{i_3}) \,\middle|\, \mathbf{y} = \boldsymbol{y}\right] \\
&\qquad - \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}\boldsymbol{y}_{i_3} \\
&= \frac{1}{\sigma^2}[\boldsymbol{\mu}_3(\boldsymbol{y})]_{i_1,i_2,i_3} + \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\mathbf{y})]_{i_1,i_2}[\boldsymbol{\mu}_1(\mathbf{y})]_{i_3} - \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}\boldsymbol{y}_{i_3}
\end{aligned} \tag{S21}$$

Substituting this back into (S19), we obtain

$$\begin{aligned}
\frac{\partial[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,i_2}}{\partial \boldsymbol{y}_{i_{k+1}}} &= \frac{1}{\sigma^2}[\boldsymbol{\mu}_3(\boldsymbol{y})]_{i_1,i_2,i_3} + \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\mathbf{y})]_{i_1,i_2}[\boldsymbol{\mu}_1(\mathbf{y})]_{i_3} - \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}\boldsymbol{y}_{i_3} \\
&\qquad - \frac{1}{\sigma^2}[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}\left([\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_3} - \boldsymbol{y}_{i_3}\right) \\
&= \frac{1}{\sigma^2}[\boldsymbol{\mu}_3(\boldsymbol{y})]_{i_1,i_2,i_3}.
\end{aligned} \tag{S22}$$

This demonstrates that

$$[\boldsymbol{\mu}_3(\boldsymbol{y})]_{i_1,i_2,i_3} = \sigma^2\frac{\partial[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,i_2}}{\partial \boldsymbol{y}_{i_{k+1}}}, \tag{S23}$$

which completes the proof for $k = 2$.

When $k \geq 3$, none of the terms in (S20) vanish, and the expression reads

$$\begin{aligned}
\frac{\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}}}{p_{\mathbf{y}}(\boldsymbol{y})} &= -\left([\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{i_2,\ldots,i_k}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_{k+1}}} + \cdots + [\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{i_1,\ldots,i_{k-1}}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k}}{\partial \boldsymbol{y}_{i_{k+1}}}\right) \\
&\quad - \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\boldsymbol{y}_{i_{k+1}} + \frac{1}{\sigma^2}\mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})\cdots(\mathbf{x}_{i_k} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_k})\mathbf{x}_{i_{k+1}} \,\middle|\, \mathbf{y} = \boldsymbol{y}\right] \\
&= -\sum_{j=1}^{d}[\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_j}}{\partial \boldsymbol{y}_{i_{k+1}}} - \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\boldsymbol{y}_{i_{k+1}} \\
&\quad + \frac{1}{\sigma^2}\mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_1})\cdots(\mathbf{x}_{i_k} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_k})\left(\mathbf{x}_{i_{k+1}} - [\boldsymbol{\mu}_1(\mathbf{y})]_{i_{k+1}} + [\boldsymbol{\mu}_1(\mathbf{y})]_{i_{k+1}}\right) \,\middle|\, \mathbf{y} = \boldsymbol{y}\right] \\
&= -\sum_{j=1}^{k}[\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_j}}{\partial \boldsymbol{y}_{i_{k+1}}} - \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\boldsymbol{y}_{i_{k+1}} + \frac{1}{\sigma^2}[\boldsymbol{\mu}_{k+1}(\boldsymbol{y})]_{i_1,\ldots,i_{k+1}} \\
&\quad + \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_{k+1}} \\
&= -\sum_{j=1}^{k}[\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_j}}{\partial \boldsymbol{y}_{i_{k+1}}} + \frac{1}{\sigma^2}[\boldsymbol{\mu}_{k+1}(\boldsymbol{y})]_{i_1,\ldots,i_{k+1}} \\
&\quad + \frac{1}{\sigma^2}[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}\left([\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_{k+1}} - \boldsymbol{y}_{i_{k+1}}\right),
\end{aligned} \tag{S24}$$

where we used the definition $\boldsymbol{\ell}_j \triangleq (i_1, \ldots, i_{j-1}, i_{j+1} \ldots, i_k)$. Substituting this expression back into (S19), we obtain

$$\frac{\partial[\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} = -\sum_{j=1}^{k}[\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j}\frac{\partial[\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_j}}{\partial \boldsymbol{y}_{i_{k+1}}} + \frac{1}{\sigma^2}[\boldsymbol{\mu}_{k+1}(\boldsymbol{y})]_{i_1,\ldots,i_{k+1}}. \tag{S25}$$

This demonstrates that

$$
\begin{aligned}
[\boldsymbol{\mu}_{k+1}(\boldsymbol{y})]_{i_1,\dots,i_{k+1}} &= \sigma^2 \frac{\partial [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\dots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} + \sigma^2 \sum_{j=1}^{k} [\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j} \frac{\partial [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_j}}{\partial \boldsymbol{y}_{i_{k+1}}} \\
&= \sigma^2 \frac{\partial [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\dots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} + \sum_{j=1}^{k} [\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j} [\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_j,i_{k+1}}, \qquad \text{(S26)}
\end{aligned}
$$

where we used (S17). This completes the proof for $k \geq 3$.

C Proof of Theorem 3

We will use the fact that for any $k \geq 1$, the posterior kth order central moment of $\boldsymbol{v}^\top \mathbf{x}$ can be written explicitly by expanding brackets as

$$
\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{v}^\top(\mathbf{x} - \boldsymbol{\mu}_1(\boldsymbol{y}))\right)^k \Big| \mathbf{y} = \boldsymbol{y}\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{d} \boldsymbol{v}_i \left[\mathbf{x} - \boldsymbol{\mu}_1(\boldsymbol{y})\right]_i\right)^k \Bigg| \mathbf{y} = \boldsymbol{y}\right] \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_k} \mathbb{E}\left[(\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}) \dots (\mathbf{x}_{i_1} - [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_k}) | \mathbf{y} = \boldsymbol{y}\right] \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_k} [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\dots,i_k}. \qquad \text{(S27)}
\end{aligned}
$$

Let us start with the second moment. From (S27), it is given by

$$
\begin{aligned}
\mu_2^{\boldsymbol{v}}(\boldsymbol{y}) &= \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \boldsymbol{v}_{i_1} \boldsymbol{v}_{i_2} [\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2} \\
&= \boldsymbol{v}^\top \boldsymbol{\mu}_2(\boldsymbol{y}) \boldsymbol{v} \\
&= \sigma^2 \boldsymbol{v}^\top \frac{\partial \boldsymbol{\mu}_1(\boldsymbol{y})}{\partial \boldsymbol{y}} \boldsymbol{v} \\
&= \sigma^2 \nabla_{\boldsymbol{y}} \left(\boldsymbol{v}^\top \boldsymbol{\mu}_1(\boldsymbol{y})\right)^\top \boldsymbol{v} \\
&= \sigma^2 D_{\boldsymbol{v}} \left(\boldsymbol{v}^\top \boldsymbol{\mu}_1(\boldsymbol{y})\right) \\
&= \sigma^2 D_{\boldsymbol{v}} \mu_1^{\boldsymbol{v}}(\boldsymbol{y}). \qquad \text{(S28)}
\end{aligned}
$$

This proves the first line of (14).

Next, we derive the third moment. From (S27), it is given by

$$
\begin{aligned}
\mu_3^{\boldsymbol{v}}(\boldsymbol{y}) &= \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \sum_{i_3=1}^{d} \boldsymbol{v}_{i_1} \boldsymbol{v}_{i_2} \boldsymbol{v}_{i_3} [\boldsymbol{\mu}_3(\boldsymbol{y})]_{i_1,i_2,i_3} \\
&= \sigma^2 \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \sum_{i_3=1}^{d} \boldsymbol{v}_{i_1} \boldsymbol{v}_{i_2} \boldsymbol{v}_{i_3} \frac{\partial [\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2}}{\partial \boldsymbol{y}_{i_3}} \\
&= \sigma^2 \sum_{i_3=1}^{d} \boldsymbol{v}_{i_3} \frac{\partial \left(\boldsymbol{v}^\top \boldsymbol{\mu}_2(\boldsymbol{y}) \boldsymbol{v}\right)}{\partial \boldsymbol{y}_{i_3}} \\
&= \sigma^2 \boldsymbol{v}^\top \nabla_{\boldsymbol{y}} \left(\boldsymbol{v}^\top \boldsymbol{\mu}_2(\boldsymbol{y}) \boldsymbol{v}\right) \\
&= \sigma^2 D_{\boldsymbol{v}} \left(\boldsymbol{v}^\top \boldsymbol{\mu}_2(\boldsymbol{y}) \boldsymbol{v}\right) \\
&= \sigma^2 D_{\boldsymbol{v}} \mu_2^{\boldsymbol{v}}(\boldsymbol{y}), \qquad \text{(S29)}
\end{aligned}
$$

where in the last line we used (S28). This proves the second line of (14).

Finally, we derive the $(k+1)$th moment for any $k \geq 3$. From (S27), it is given by

$$
\begin{aligned}
\mu^{\boldsymbol{v}}_{k+1}(\boldsymbol{y}) &= \sum_{i_1=1}^{d} \cdots \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_{k+1}} [\boldsymbol{\mu}_{k+1}(\boldsymbol{y})]_{i_1,\dots,i_{k+1}} \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_{k+1}} \left(\sigma^2 \frac{\partial [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\dots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} + \sum_{j=1}^{k} [\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j} [\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_j, i_{k+1}} \right) \\
&= \sigma^2 \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_{i_{k+1}} \frac{\partial}{\partial \boldsymbol{y}_{i_{k+1}}} \left(\sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_k} [\boldsymbol{\mu}_k(\boldsymbol{y})]_{i_1,\dots,i_k} \right) + \\
&\quad \sum_{j=1}^{k} \left(\sum_{i_1=1}^{d} \cdots \sum_{i_{j-1}=1}^{d} \sum_{i_{j+1}=1}^{d} \cdots \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_{i_1} \dots \boldsymbol{v}_{i_{j-1}} \boldsymbol{v}_{i_{j+1}} \dots \boldsymbol{v}_{i_k} [\boldsymbol{\mu}_{k-1}(\boldsymbol{y})]_{\boldsymbol{\ell}_j} \sum_{i_j=1}^{d} \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_j \boldsymbol{v}_{i_{k+1}} [\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_j, i_{k+1}} \right) \\
&= \sigma^2 \sum_{i_{k+1}=1}^{d} \boldsymbol{v}_{i_{k+1}} \frac{\partial \mu^{\boldsymbol{v}}_k(\boldsymbol{y})}{\partial \boldsymbol{y}_{i_{k+1}}} + \sum_{j=1}^{k} \mu^{\boldsymbol{v}}_{k-1}(\boldsymbol{y}) \mu^{\boldsymbol{v}}_2(\boldsymbol{y}) \\
&= \sigma^2 \boldsymbol{v}^\top \nabla_{\boldsymbol{y}} \mu^{\boldsymbol{v}}_k(\boldsymbol{y}) + k \mu^{\boldsymbol{v}}_{k-1}(\boldsymbol{y}) \mu^{\boldsymbol{v}}_2(\boldsymbol{y}) \\
&= \sigma^2 D_{\boldsymbol{v}} \mu^{\boldsymbol{v}}_k(\boldsymbol{y}) + k \mu^{\boldsymbol{v}}_{k-1}(\boldsymbol{y}) \mu^{\boldsymbol{v}}_2(\boldsymbol{y}),
\end{aligned} \tag{S30}
$$

where in the second line we used (10). This completes the proof of the third line of (14).

D Related work: estimation of higher order scores by denoising

The work most related to ours is that of Meng et al. (2021). Here, we present their results while translating to our notation. Given a probability density $p_{\mathbf{y}}$ over $\mathbb{R}^d$, they define the kth order score $\mathbf{s}_k(\boldsymbol{y})$ as the tensor whose entry at multi-index $(i_1, i_2, ..., i_k)$ is

$$
[\mathbf{s}_k(\boldsymbol{y})]_{i_1,i_2,\dots,i_k} \triangleq \frac{\partial^k}{\partial \boldsymbol{y}_{i_1} \partial \boldsymbol{y}_{i_2} \dots \partial \boldsymbol{y}_{i_k}} \log p_{\mathbf{y}}(\boldsymbol{y}), \tag{S31}
$$

for every $i_1, \dots, i_k \in \{1, \dots, d\}^k$. Using our notation, and under the assumption (5) that $\mathbf{y}$ is a noisy version of $\mathbf{x} \sim p_{\mathbf{x}}$, the denoising score matching method estimates the first-order score $\mathbf{s}_1(\boldsymbol{y})$, which is simply the gradient of the log-probability, $\nabla_{\mathbf{y}} \log p_{\mathbf{y}}(\boldsymbol{y})$. This is done by using Tweedie's formula, which links $\mathbf{s}_1$ with the first posterior moment (the MSE-optimal denoiser) as

$$
\boldsymbol{\mu}_1(\boldsymbol{y}) = \mathbb{E}[\mathbf{x}|\mathbf{y} = \boldsymbol{y}] = \boldsymbol{y} + \sigma^2 \mathbf{s}_1(\boldsymbol{y}). \tag{S32}
$$

As noted by Meng et al. (2021), a similar relation links the second-order score with the second posterior moment (*i.e.*, the posterior covariance) as

$$
\boldsymbol{\mu}_2(\boldsymbol{y}) = \mathrm{Cov}(\mathbf{x}|\mathbf{y} = \boldsymbol{y}) = \sigma^4 \mathbf{s}_2(\boldsymbol{y}) + \sigma^2 I. \tag{S33}
$$

Note from (S31) that $\mathbf{s}_2(\boldsymbol{y})$ is the Hessian of the log-probability, $\nabla^2_{\mathbf{y}} \log p_{\mathbf{y}}(\boldsymbol{y})$, or equivalently the Jacobian of the gradient of the log probability, $\frac{\partial}{\partial \boldsymbol{y}} \nabla_{\mathbf{y}} \log p_{\mathbf{y}}(\boldsymbol{y})$. And since we have from (S32) that $\nabla_{\boldsymbol{y}} \log p_{\mathbf{y}}(\boldsymbol{y}) = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1(\boldsymbol{y}) - \boldsymbol{y})$, Eq. (S33) can be equivalently written as

$$
[\boldsymbol{\mu}_2(\boldsymbol{y})]_{i_1,i_2} = \sigma^4 \frac{\partial}{\partial \boldsymbol{y}_{i_2}} \left[\frac{\mu_1(\boldsymbol{y}) - \boldsymbol{y}}{\sigma^2} \right]_{i_1} + \sigma^2 I = \sigma^2 \frac{\partial [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_1}}{\partial \boldsymbol{y}_{i_2}}. \tag{S34}
$$

This illustrates that the second-order formula of Meng et al. (2021) is equivalent to (10).

Moving on to higher-order moments, following our notations, Lemma 1 of Meng et al. (2021) states that

$$
\mathbb{E}[\otimes^{k+1} \mathbf{x}|\mathbf{y} = \boldsymbol{y}] = \sigma^2 \frac{\partial}{\partial \boldsymbol{y}} \mathbb{E}[\otimes^k \mathbf{x}|\mathbf{y} = \boldsymbol{y}] + \sigma^2 \mathbb{E}[\otimes^k \mathbf{x}|\mathbf{y} = \boldsymbol{y}] \otimes \left(\mathbf{s}_1(\boldsymbol{y}) + \frac{\boldsymbol{y}}{\sigma^2} \right), \quad \forall k \geq 1, \tag{S35}
$$

where $\otimes^{k+1}\mathbf{x} \in \mathbb{R}^{d^k}$ denotes k-fold tensor multiplication. This lemma is used in Theorem 3 of Meng et al. (2021), to derive a recursion relating higher-order moments and scores. Substituting (S32), this relation can be written as

$$\mathbb{E}[\otimes^{k+1}\mathbf{x}|\mathbf{y}=\boldsymbol{y}] = \sigma^2 \frac{\partial}{\partial \boldsymbol{y}} \mathbb{E}[\otimes^{k}\mathbf{x}|\mathbf{y}=\boldsymbol{y}] + \mathbb{E}[\otimes^{k}\mathbf{x}|\mathbf{y}=\boldsymbol{y}] \otimes \boldsymbol{\mu}_1(\boldsymbol{y}), \quad \forall k \geq 1. \tag{S36}$$

Denoting the non-central posterior moment of order k by $\boldsymbol{m}_k(\boldsymbol{y})$, Eq. (S36) can be written compactly as

$$\boldsymbol{m}_{k+1}(\boldsymbol{y}) = \sigma^2 \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{m}_k(\boldsymbol{y}) + \boldsymbol{m}_k(\boldsymbol{y}) \otimes \boldsymbol{\mu}_1(\boldsymbol{y}), \quad \forall k \geq 1. \tag{S37}$$

Writing out the elements of $\boldsymbol{m}_{k+1}(\boldsymbol{y})$ explicitly, this relation reads

$$[\boldsymbol{m}_{k+1}(\boldsymbol{y})]_{i_1,\ldots,i_{k+1}} = \sigma^2 \frac{\partial [\boldsymbol{m}_k(\boldsymbol{y})]_{i_1,\ldots,i_k}}{\partial \boldsymbol{y}_{i_{k+1}}} + [\boldsymbol{m}_k(\boldsymbol{y})]_{i_1,\ldots,i_k} [\boldsymbol{\mu}_1(\boldsymbol{y})]_{i_{k+1}}, \quad \forall k \geq 1. \tag{S38}$$

It is interesting to compare this expression with the recursion for the central moments in Theorem 2. We see that the non-central moments satisfy a sort of one-step recursion (if we disregard the dependence on $\boldsymbol{\mu}_1$), in the sense that $\boldsymbol{m}_{k+1}$ depends only on $\boldsymbol{m}_k$. In contrast, as can be seen in Theorem 2, the central moments satisfy a sort of two-step recursion (if we disregard the dependence on $\boldsymbol{\mu}_2$), in the sense that $\boldsymbol{\mu}_{k+1}(\boldsymbol{y})$ depends on both $\boldsymbol{\mu}_k(\boldsymbol{y})$ and $\boldsymbol{\mu}_{k-1}(\boldsymbol{y})$.

E POSTERIOR DISTRIBUTION FOR A GAUSSIAN MIXTURE PRIOR

In Fig. 1 and Fig. 3, we demonstrated our approach on one-dimensional and two-dimensional Gaussian mixtures, respectively. In both cases, we showed plots of the marginal posterior distribution in the direction of the first posterior principal component, as well as the posterior mean for a particular noisy input sample. Those simulations relied on the closed-form expressions of the posterior distribution and the marginal posterior distribution along some direction for a Gaussian mixture prior. In addition, Fig. 1 and Fig. 3 also contain the maximum entropy distribution estimated using our method, which uses the numerical derivatives of the posterior mean. Here as well we used the numerical derivatives of the posterior mean function, which we computed in closed-from. We now present these closed-form expressions for completeness.

Suppose $p_{\mathbf{x}}$ is a mixture of L Gaussians,

$$p_{\mathbf{x}}(\boldsymbol{x}) = \sum_{\ell=1}^{L} \pi_\ell \mathcal{N}(\boldsymbol{x}; m_\ell, \Sigma_\ell). \tag{S39}$$

Let c be a random variable taking values in $\{1,\ldots,L\}$ with probabilities $\pi_1,\ldots,\pi_L$. Then we can think of $\mathbf{x}$ as drawn from the ℓth Gaussian conditioned on the event that $\mathrm{c}=\ell$. Therefore,

$$\begin{aligned}
p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) &= \sum_{\ell=1}^{L} p_{\mathbf{x}|\mathbf{y},\mathrm{c}}(\boldsymbol{x}|\boldsymbol{y},\ell) p_{\mathrm{c}|\mathbf{y}}(\ell|\boldsymbol{y}) \\
&= \sum_{\ell=1}^{L} p_{\mathbf{x}|\mathbf{y},\mathrm{c}}(\boldsymbol{x}|\boldsymbol{y},\ell) \frac{p_{\mathbf{y}|\mathrm{c}}(\boldsymbol{y}|\ell) p_{\mathrm{c}}(\ell)}{p_{\mathbf{y}}(\boldsymbol{y})} \\
&= \sum_{\ell=1}^{L} \mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{m}}_\ell, \bar{\Sigma}_\ell) \frac{\rho_\ell \pi_\ell}{\sum_{\ell'=1}^{L} \rho_{\ell'} \pi_{\ell'}},
\end{aligned} \tag{S40}$$

where we denoted

$$\begin{aligned}
\rho_i &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{m}_i, \Sigma_i + \sigma^2 \boldsymbol{I}), \\
\bar{\boldsymbol{m}}_i &= \Sigma_i (\Sigma_i + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{m}_i) + \boldsymbol{m}_i, \\
\bar{\Sigma}_i &= \Sigma_i - \Sigma_i (\Sigma_i + \sigma^2 \boldsymbol{I})^{-1} \Sigma_i.
\end{aligned} \tag{S41}$$

As for the marginal posterior distribution along some direction $\boldsymbol{v}$, it is easy to show that

$$
\begin{aligned}
p_{\boldsymbol{v}^\top \mathbf{x}|\mathbf{y}}(\alpha|\boldsymbol{y}) &= \sum_{\ell=1}^{L} p_{\boldsymbol{v}^\top \mathbf{x}|\mathbf{y},\mathrm{c}}(\alpha|\boldsymbol{y},\ell) p_{\mathrm{c}|\mathbf{y}}(\ell|\boldsymbol{y}) \\
&= \sum_{\ell=1}^{L} p_{\boldsymbol{v}^\top \mathbf{x}|\mathbf{y},\mathrm{c}}(\alpha|\boldsymbol{y},\ell) \frac{p_{\mathbf{y}|\mathrm{c}}(\boldsymbol{y}|\ell) p_{\mathrm{c}}(\ell)}{p_{\mathbf{y}}(\boldsymbol{y})} \\
&= \sum_{\ell=1}^{L} \mathcal{N}(\alpha; \boldsymbol{v}^\top \bar{\boldsymbol{m}}_\ell, \boldsymbol{v}^\top \bar{\Sigma}_\ell \boldsymbol{v}) \frac{\rho_\ell \pi_\ell}{\sum_{\ell'=1}^{L} \rho_{\ell'} \pi_{\ell'}}.
\end{aligned} \tag{S42}
$$

F Proof of Corollary 1

We start by reminding the reader of (4) :

$$
\begin{aligned}
\mu_2(y) &= \sigma^2 \mu_1'(y), \\
\mu_3(y) &= \sigma^2 \mu_2'(y), \\
\mu_{k+1}(y) &= \sigma^2 \mu_k'(y) + k\mu_{k-1}(y)\mu_2(y), \quad k \geq 3.
\end{aligned}
$$

We will prove by complete induction that

$$
\mu_k^{(m)} = 0 \quad \text{for all } \; k \geq 2 \; \text{ and } \; m \geq 1. \tag{S43}
$$

Base Note that since for any $m \geq 2$ we have $\mu_1^{(m)}(y^*) = 0$, for any $m \geq 1$ we have

$$
\begin{aligned}
\mu_2^{(m)}(y^*) &= \sigma^2 \mu_1^{(m+1)}(y^*) \\
&= 0 \\
\mu_3^{(m)}(y^*) &= \sigma^2 \mu_2^{(m+1)}(y^*) \\
&= \sigma^4 \mu_1^{(m+2)}(y^*) \\
&= 0 \\
\mu_4^{(m)}(y^*) &= \sigma^2 \mu_3^{(m+1)}(y^*) + 3 \, \frac{\partial^m}{\partial y^m} \left(\mu_2^2(y)\right)\Big|_{y=y^*} \\
&\overset{(1)}{=} \sigma^2 \mu_3^{(m+1)}(y^*) + 3 \sum_{l=0}^{m} \binom{m}{l} \mu_2^{(m-l)}(y^*) \mu_2^{(l)}(y^*) \\
&= \sigma^2 \mu_3^{(m+1)}(y^*) + 3 \left(\mu_2^{(m)}(y^*)\mu_2(y^*) + \cdots + \mu_2(y^*)\mu_2^{(m)}(y^*)\right) \\
&= \sigma^2 \mu_3^{(m+1)}(y^*) \\
&= 0,
\end{aligned} \tag{S44}
$$

where (1) results from the general Leibniz rule.

Induction Assume that $\mu_n^{(m)}(y^*) = 0$ for all $4 \leq n < k+1$ and $m \geq 1$. Then,

$$
\begin{aligned}
\mu_{k+1}^{(m)}(y^*) &= \frac{\partial^m}{\partial y^m} \left(\sigma^2 \mu_k'(y) + k\mu_{k-1}(y)\mu_2(y)\right)\Big|_{y=y*} \\
&= \sigma^2 \mu_k^{(m+1)}(y^*) + k \, \frac{\partial^m}{\partial y^m} \left(\mu_{k-1}(y)\mu_2(y)\right)\Big|_{y=y^*} \\
&\overset{(1)}{=} \sigma^2 \mu_k^{(m+1)}(y^*) + k \sum_{l=0}^{m} \binom{m}{l} \mu_{k-1}^{(m-l)}(y^*) \mu_2^{(l)}(y^*) \\
&= \sigma^2 \mu_k^{(m+1)}(y^*) + k\mu_{k-1}^{(m)}(y^*)\mu_2(y^*) + ... + k\mu_{k-1}(y^*)\mu_2^{(m)}(y^*) \\
&\overset{(2)}{=} 0,
\end{aligned} \tag{S45}
$$

where for (1) the general Leibniz rule was used again, and in (2) we used our induction assumption. This concludes the induction.

Using (S43) we therefore obtain for all $k \geq 3$ that

$$\begin{aligned} \mu_{k+1}(y^*) &= k\mu_{k-1}(y^*)\mu_2(y^*), \\ &= k(k-2)\mu_2^2(y^*)\mu_{k-3}(y^*) \\ &= k(k-2)(k-4)\mu_2^3(y^*)\mu_{k-5}(y^*) \\ &= \ldots \\ &= \begin{cases} k!!\mu_2^{\frac{k+1}{2}}(y^*) & k \text{ is odd,} \\ 0 & k \text{ is even.} \end{cases} \end{aligned} \tag{S46}$$

Since $\mu_3(y^*) = \sigma^2\mu_2(y^*) = 0$ as well, the posterior moments are the same as those of a Gaussian distribution. Indeed, the central moments of a random variable $\mathrm{z} \sim \mathcal{N}(\mathbb{E}[\mathrm{z}], \sigma^2)$ are given by

$$\mathbb{E}[(\mathrm{z} - \mathbb{E}[\mathrm{z}])^d] = \begin{cases} \sigma^d(d-1)!! & d \text{ is even,} \\ 0 & d \text{ is odd.} \end{cases} \tag{S47}$$

To conclude the proof, all that remains to show is moment-determinacy (*i.e.*, that the sequence of moments uniquely determines the distribution). This is the case, since the moments of a Gaussian distribution are trivially verified to satisfy *e.g.*, Condition $(h6)$ of (Lin, 2017). This implies that the posterior is moment-determinate, and is Gaussian.

G EXPERIMENTAL DETAILS

Algorithm 1 requires three hyper-parameters as input. The first is the small constant c, which is used for the linear approximation in (15). The second is N, which is the number of principal components we seek. The last is K, which is the number of iterations to preform. In all our experiments we used $c = 10^{-5}$ and $N = 3$. For the N2V experiments we used $K = 100$ while for the rest we used $K = 50$.

Figure S1 depicts the convergence of the subspace iteration method for two different domains. For each noisy image and patch for which we find the principal components (marked in red), the plot to the right shows the convergence of the first $N = 3$ principal components. Specifically, for each principal component $\boldsymbol{v}_i$, we calculate its inner product with the same principal component in the previous iteration. As the graph shows, $K = 50$ iterations suffice for convergence.

H THE IMPACT OF THE JACOBIAN-VECTOR DOT-PRODUCT LINEAR APPROXIMATION

As described in Sec 4.1, Alg. 1 calls for calculating the Jacobian-vector dot-product of the denoiser. While for neural denoisers this calculation can be done via automatic differentiation, we propose using a linear approximation instead (See Eq. (15)). This can reduce the computational burden, while retaining high-accuracy in the computed eigenvectors. For example, in an experiment using SwinIR and $\sigma = 50$, the cosine similarity between the principal components computed with the approximation and those computed with automatic differentiation typically reaches around 0.97 at the $50th$ iteration. However, in terms of computational burden, the differences can sometimes be dramatic. For example, with the SwinIR model, when calculating one eigenvector for a patch of size 80×92, the memory footprint using automatic differentiation reaches 12GB, while using the linear approximation method it only reaches 2GB. These differences will increase for running on larger images. A visual comparison of the resulting principal component can be found in Fig. S2.

I THE IMPACT OF ESTIMATING σ

Our theoretical analysis is developed for non-blind denoising, and accordingly, most of our experiments conform to this setting. These include the experiments on faces (Fig. 2 and Fig. S6), on

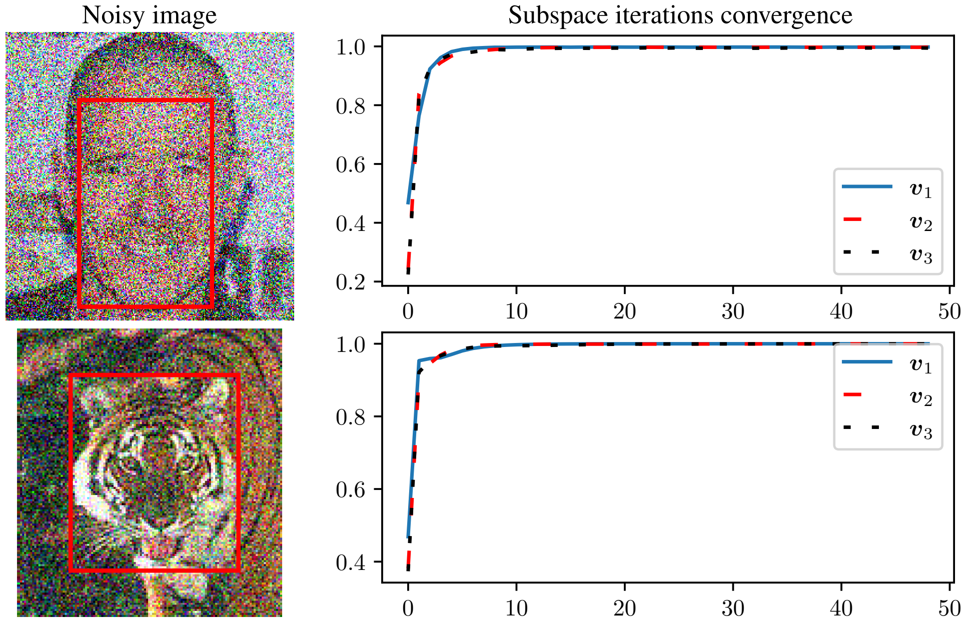

Figure S1: **Convergence of the subspace iteration method**. In each row one noisy image is shown with a red patch marking the region for which the posterior principal components are calculated. To the right, we plot for each of the first 3 principal components the inner product between the principal component in consecutive iterations. As the graph shows, $K = 50$ iterations suffice to guarantee convergence in those domains.

MNIST digits (Fig. 4), on natural images (top part of Fig. 5, S4 and S5), and the toy problem of Fig. 3. Namely, in all those experiments the noise level σ was assumed known.

Nevertheless, we show empirically that our method can also work well in the blind setting. This is the case in the real microscopy images (bottom part of Fig. 5). In this experiment, we estimated σ using the naive formula $\hat{\sigma}^2 = \frac{1}{d}\|\boldsymbol{\mu}_1(\boldsymbol{y}) - \boldsymbol{y}\|^2$, where $\boldsymbol{\mu}_1(\boldsymbol{y})$ is the (blind) N2V denoiser. It is certainly possible to employ more advanced noise-level estimation methods in order to obtain an even more accurate estimate for σ. Indeed, noise-level estimation, particularly for white Gaussian noise, has been heavily researched, and as of today state-of-the-art methods reach very-high precision (Chen et al., 2015; Khmag et al., 2018; Kokil & Pratap, 2021; Liu & Lin, 2012; Liu et al., 2013). For example, when the real σ equals 10, the error in estimating sigma is around 0.05 (see *e.g.*, Chen et al. (2015)). However, we find that even with the naive method described above, we get quite accurate results. Particularly, the impact of small inaccuracies in σ on our uncertainty estimation turn out to be very small. To illustrate this, we applied our method with a SwinIR model that was trained for $\sigma = 50$, on images with noise levels of $\sigma = 47.5, 52.5$. This accounts for 5% errors in σ, that are significantly higher than typical 0.5% errors of good noise level estimation techniques. Despite the inaccuracies in σ, the eigenvectors produced using our method are quite similar, as can be seen in Fig. S3.

J Additional results

Figures S4 and S5 provide additional results on test images from the McMaster (Zhang et al., 2011) dataset and images from ImageNet (Deng et al., 2009). In the anonymized repository, we attach a video showing more examples on face images, demonstrating different semantic principal components.

J.1 Polynomial fitting examples

As discussed briefly in Sec. 5, we experimented with fitting a polynomial to the function $f(\alpha) = \boldsymbol{v}^\top \boldsymbol{\mu}_1(\boldsymbol{y} + \alpha \boldsymbol{v})$, and using the derivatives of the polynomial at $\alpha = 0$ instead of using numerical

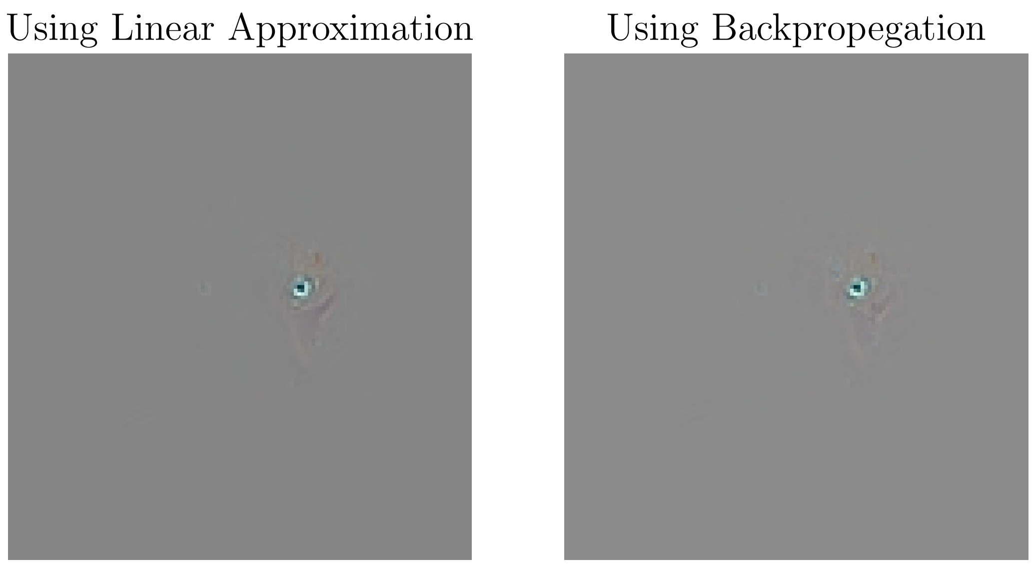

Figure S2: **The impact of the linear approximation on the calculated principal component.** The first principal component calculated with SwinIR and $\sigma = 50$, using the linear approximation in Eq. (15), and using automatic differentiation (Backpropegation). Both methods achieve similar results, with a cosine similarity of 0.96 over 50 iterations. However, the linear approximation methods uses drastically less memory.

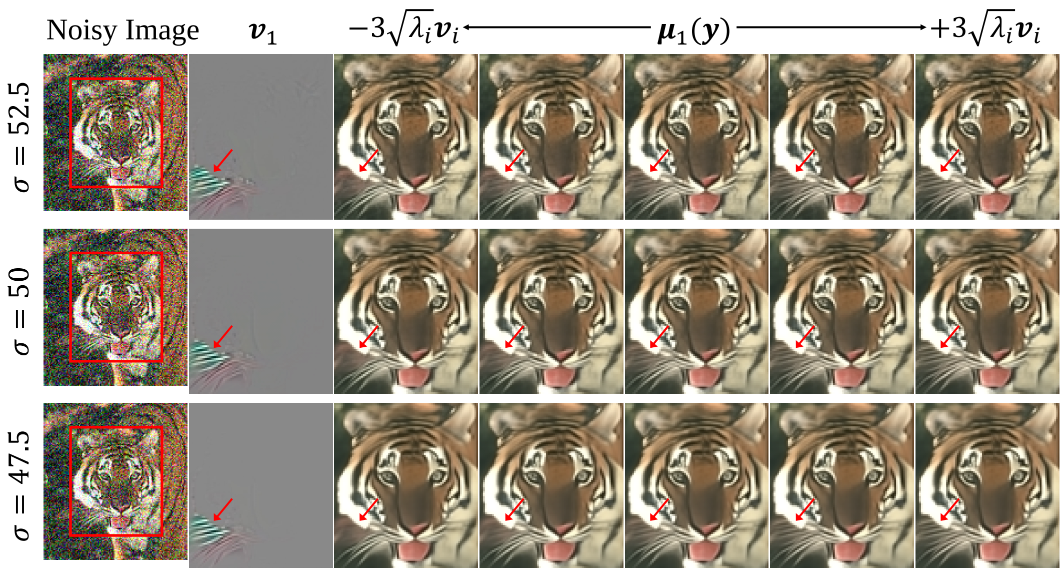

Figure S3: **The effect of small inaccuracies in σ on uncertainty estimation.** The first principal component calculated using SwinIR, for an assumed $\sigma = 50$, for three different actual noise levels in the image.

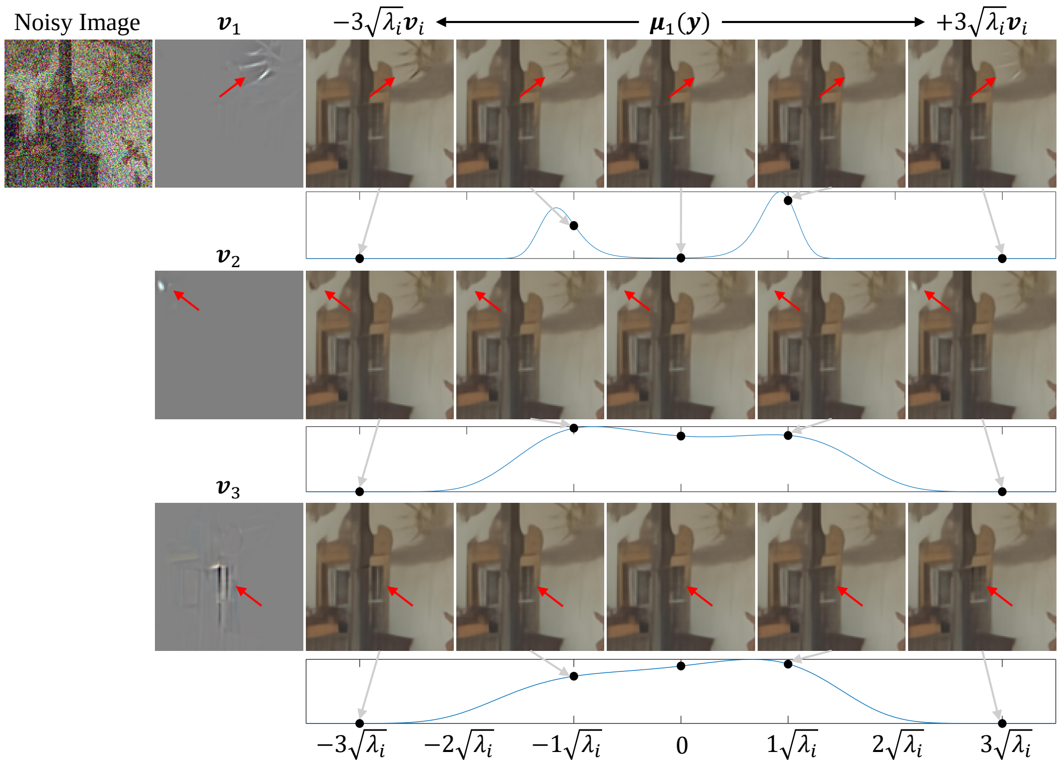

Figure S4: **Additional examples on natural images using SwinIR** (Liang et al., 2021). In each row, one of the first three PCs corresponding to the noisy image is shown on the left. On the right, images along the PC are shown above the marginal posterior distribution estimated for this direction. The principal components reveal uncertainty in delicate parts of the wall-painting, such as the thin rays of the sun, or the existence of mullions in the windows.

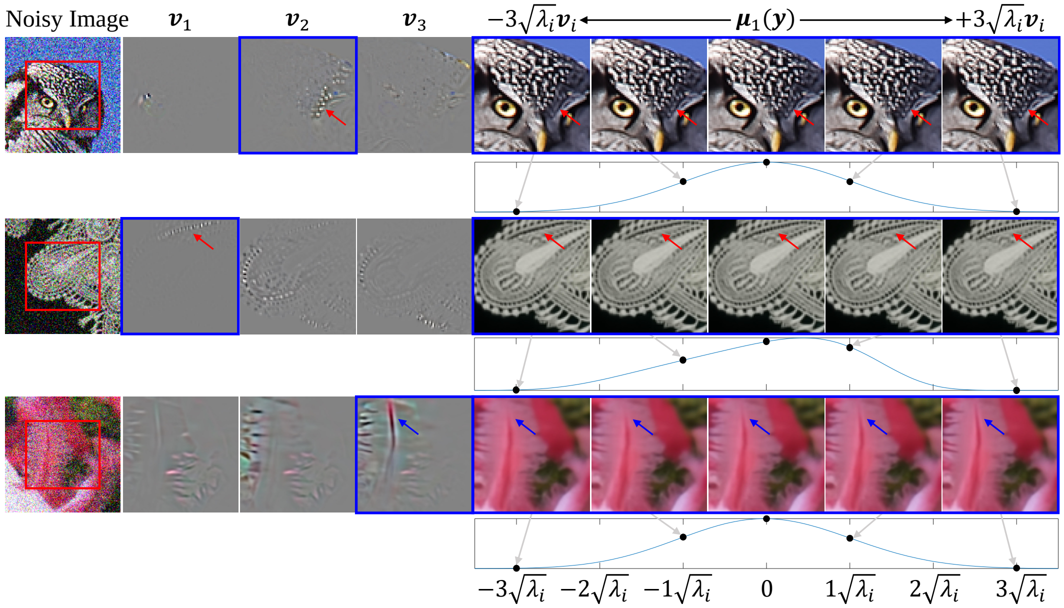

Figure S5: **Additional examples on natural images using SwinIR** (Liang et al., 2021). In each row, the first three PCs corresponding to the noisy image are shown on the left, and one is marked in blue. On the right, images along the marked PC are shown above the marginal posterior distribution estimated for this direction. The principal components catch semantic directions such as the pattern on the owl's feathers, the embroidery pattern, or the length of the Axolotl's gills.

derivatives of $f(\alpha)$ itself at $\alpha = 0$. Here, we provide the results of an experiment where we fit a polynomial of degree six over the range $\left[-\sqrt{\lambda_i}, \sqrt{\lambda_i}\right]$ for the ith principal component. As can be seen in Fig. S6, the low degree polynomial smooths the directional posterior mean function, thus leading to a smooth Gaussian-like marginal posterior distribution estimate. Presumably, these posterior estimates are smoother than the true posterior.

Figure S6: **Additional examples on face images, using a polynomial fit marginal distribution estimate**. In each row, the first three PCs corresponding to the noisy image are shown on the left, and one is marked in blue. On the right, images along the marked PC are shown above the marginal posterior distribution estimated for this direction. The principal components highlight meaningful uncertainty, such as eyes shape or the existence of wrinkles. Note as an example in the first row how the optimal-MSE restoration is the mean of the more probable mode, depicting no hair on the forehead, and the distribution's tail, yielding the less-probable semi-translucent hair.

REFERENCES

Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 477–485, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Asem Khmag, Abd Rahman Ramli, SA Al-Haddad, and Noraziahtulhidayu Kamarudin. Natural image noise level estimation based on local statistics for blind noise reduction. *The Visual Computer: International Journal of Computer Graphics*, 34(4):575–587, 2018.

Priyanka Kokil and Turimerla Pratap. Additive white gaussian noise level estimation for natural images using linear scale-space features. *Circuits, Systems, and Signal Processing*, 40(1):353–374, 2021.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Gwo Dong Lin. Recent developments on the moment problem. *Journal of Statistical Distributions and Applications*, 4(1):5, 2017.

Wei Liu and Weisi Lin. Additive white gaussian noise level estimation in svd domain for images. *IEEE Transactions on Image processing*, 22(3):872–883, 2012.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE transactions on image processing*, 22(12):5226–5237, 2013.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25359–25369. Curran Associates, Inc., 2021.

Koichi Miyasawa et al. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic Imaging*, 20(2):023016, 2011.