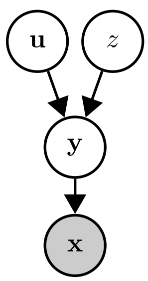
Derivations for the GSM

Máté Lengyel

February 11, 2011

1 Definition



Following Wainwright and Simoncelli (2000), we define the GSM as the following generative model:[1]

$$\mathbf{u} \sim \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{C}) \tag{1}$$

$$z \sim \mathrm{Gamma}(k, \theta) \tag{2}$$

$$\mathbf{y} = z\,\mathbf{u} \tag{3}$$

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}\mathbf{y}, \sigma_{\mathrm{x}}^2 \mathbf{I}\right) \tag{4}$$

$$\mathrm{P}(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} \mathrm{P}(\mathbf{x}_t) \tag{5}$$

[1] This is in fact a slight extension of the original model because Wainwright and Simoncelli (2000) only defined it down to the level of filter coefficient, $\mathbf{y}$ (implicitly assuming that deterministic estimates at that level suffice), while we define the model all the way down to the level of pixels, $\mathbf{x}$.

2 Some preliminaries

$$\begin{aligned}
\mathrm{P}(\mathbf{x}|\mathbf{u},z) &= \mathcal{N}\big(\mathbf{x};z\,\mathbf{A}\mathbf{u},\sigma_x^2\mathbf{I}\big) && (6)\\
&= \mathcal{N}\big(z\,\mathbf{A}\mathbf{u};\mathbf{x},\sigma_{\mathrm{x}}^2\mathbf{I}\big) && (7)\\
&= b\cdot\mathcal{N}(\mathbf{u};\mathbf{m},\mathbf{D}) \quad \text{for } N_{\mathrm{u}}\le N_{\mathrm{x}} && (8)
\end{aligned}$$

with

$$b = \frac{\sqrt{|2\pi\mathbf{D}|}}{\sqrt{|2\pi\sigma_{\mathrm{x}}^2\mathbf{I}|}}\,e^{-\frac{1}{2}\left(\frac{1}{\sigma_{\mathrm{x}}^2}\mathbf{x}^\mathsf{T}\mathbf{x}-\mathbf{m}^\mathsf{T}\mathbf{D}^{-1}\mathbf{m}\right)} \qquad = z^{-N_{\mathrm{u}}}\underbrace{\frac{\sqrt{\left|2\pi\sigma_{\mathrm{x}}^2\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\right|}}{\sqrt{|2\pi\sigma_{\mathrm{x}}^2\mathbf{I}|}}\,e^{-\frac{1}{2\sigma_{\mathrm{x}}^2}\mathbf{x}^\mathsf{T}\left[\mathbf{I}-\mathbf{A}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\right]\mathbf{x}}}_{\text{does not depend on }\mathbf{u}\text{ or }z} \tag{9}$$

$$\mathbf{m} = \mathbf{D}\,\left(z\mathbf{A}\right)^\mathsf{T}\,\frac{1}{\sigma_{\mathrm{x}}^2}\,\mathbf{x} \qquad = \frac{1}{z}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\,\mathbf{A}^\mathsf{T}\,\mathbf{x} \tag{10}$$

$$\mathbf{D}^{-1} = z^2\,\mathbf{A}^\mathsf{T}\left(\sigma_{\mathrm{x}}^2\mathbf{I}\right)^{-1}\mathbf{A} \qquad = \frac{z^2}{\sigma_{\mathrm{x}}^2}\,\mathbf{A}^\mathsf{T}\mathbf{A} \tag{11}$$

Technically, this only works for the (under)complete case (because $\mathbf{A}^\mathsf{T}\mathbf{A}$ needs to be invertable), but as we shall see later this is not going to be a problem.

3 Inference

$$\mathrm{P}(\mathbf{y},\mathbf{u},z|\mathbf{x}) = \mathrm{P}(\mathbf{y}|\mathbf{u},z,\mathbf{x})\;\mathrm{P}(\mathbf{u}|z,\mathbf{x})\;\mathrm{P}(z|\mathbf{x}) \tag{12}$$

3.1 Inferring u

$$\begin{aligned}
\mathrm{P}(\mathbf{u}|z,\mathbf{x}) &\propto \mathrm{P}(\mathbf{u})\;\mathrm{P}(\mathbf{x}|\mathbf{u},z) && (13)\\
&\propto \mathcal{N}(\mathbf{u};\mathbf{0},\mathbf{C})\;\mathcal{N}(\mathbf{u};\mathbf{m},\mathbf{D}) && (14)\\
&= \mathcal{N}(\mathbf{u};\boldsymbol{\mu}(z,\mathbf{x})\,,\boldsymbol{\Sigma}(z)) && (15)
\end{aligned}$$

with

$$\boldsymbol{\mu}(z,\mathbf{x}) = \boldsymbol{\Sigma}(z)\left(\mathbf{C}^{-1}\mathbf{0}+\mathbf{D}^{-1}\mathbf{m}\right) \qquad = \frac{z}{\sigma_{\mathrm{x}}^2}\,\boldsymbol{\Sigma}(z)\,\mathbf{A}^\mathsf{T}\mathbf{x} \tag{16}$$

$$\boldsymbol{\Sigma}(z) = \left(\mathbf{C}^{-1}+\mathbf{D}^{-1}\right)^{-1} \qquad = \left(\mathbf{C}^{-1}+\frac{z^2}{\sigma_{\mathrm{x}}^2}\,\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1} \tag{17}$$

So, as we see, both $\boldsymbol{\mu}(z, \mathbf{x})$ and $\boldsymbol{\Sigma}(z)$ are well-defined even in the overcomplete case.

We will also need the marginal posterior:

$$\mathrm{P}(\mathbf{u}|\mathbf{x}) = \sum_z \mathrm{P}(z|\mathbf{x})\ \mathrm{P}(\mathbf{u}|z, \mathbf{x}) \tag{18}$$

$$= \sum_z \mathrm{P}(z|\mathbf{x})\ \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}(z, \mathbf{x})\,, \boldsymbol{\Sigma}(z)) \tag{19}$$

3.2 Inferring $\mathbf{y}$

It follows trivially from the above:

$$\mathrm{P}(\mathbf{y}|\mathbf{u}, z, \mathbf{x}) = \delta(\mathbf{y} - z\,\mathbf{u}) \tag{20}$$

$$\mathrm{P}(\mathbf{y}|z, \mathbf{x}) = \mathcal{N}\big(\mathbf{y}; z\,\boldsymbol{\mu}(z, \mathbf{x})\,, z^2\,\boldsymbol{\Sigma}(z)\big) \tag{21}$$

$$\mathrm{P}(\mathbf{y}|\mathbf{x}) = \sum_z \mathrm{P}(z|\mathbf{x})\ \mathcal{N}\big(\mathbf{y}; z\,\boldsymbol{\mu}(z, \mathbf{x})\,, z^2\,\boldsymbol{\Sigma}(z)\big) \tag{22}$$

3.3 Inferring z

$$\mathrm{P}(z|\mathbf{x}) \propto \mathrm{P}(z)\ \mathrm{P}(\mathbf{x}|z) \tag{23}$$

Rather than deriving $\mathrm{P}(\mathbf{x}|z)$, the (marginal) likelihood of z, through lengthy algebraic manipulations, we build on the following intuition. One can think of $\mathbf{x}$ (given z) as a deterministically scaled (by $z\,\mathbf{A}$) version of $\mathbf{u}$ (a multivariate Gaussian random variable with known mean, $\mathbf{0}$, and covariance, $\mathbf{C}$) plus a (multivariate) Gaussian noise term (with $\mathbf{0}$ mean and $\sigma_{\mathrm{x}}^2\,\mathbf{I}$ covariance). This insight yields a simple form for the probability of $\mathbf{x}$ given z, which is just the likelihood we need (see also Bishop (2006, p. 93)):

$$\mathrm{P}(\mathbf{x}|z) = \mathcal{N}\big(\mathbf{x};\, \mathbf{0}, \sigma_{\mathrm{x}}^2\,\mathbf{I} + z^2\,\mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathsf{T}}\big) \tag{24}$$

4 Learning

The objective of learning is to maximise the likelihood of the parameters, $\vartheta = \left\{\sigma_{\mathrm{x}}^2, \mathbf{A}, \mathbf{C}, k, \theta\right\}$:

$$\mathrm{P}(\mathbf{x}_{1:T}|\vartheta) = \prod_t \mathrm{P}(\mathbf{x}_t|\vartheta) = \prod_t \int \mathrm{d}z\, \mathrm{P}(z|k,\theta)\ \mathrm{P}\big(\mathbf{x}_t|z, \sigma_{\mathrm{x}}^2, \mathbf{A}, \mathbf{C}\big) \tag{25}$$

$$= \prod_t \int \mathrm{d}z\, \mathrm{P}(z|k,\theta)\ \mathcal{N}\big(\mathbf{x}_t;\, \mathbf{0}, \sigma_{\mathrm{x}}^2\, \mathbf{I} + z^2\, \mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathsf{T}}\big) \tag{26}$$

where in the last step we used Eq. 24.

Equation 26 reveals important invariances in the model. Namely, the likelihood (or predictive density) only depends on $\mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathsf{T}}$ and is thus invariant under the following reparametrisation: $\mathbf{C} \rightarrow \mathbf{C}'$, $\mathbf{A} \rightarrow \mathbf{A}\,\mathbf{C}^{\frac{1}{2}}\,\mathbf{C}'^{-\frac{1}{2}}$. Thus, it does not make sense to learn *both* $\mathbf{C}$ and $\mathbf{A}$. Likewise, the scale of z (parametrised by θ in the case of the Gamma prior we are using) can also be folded into either $\mathbf{A}$ or $\mathbf{C}$, since z simply multiplies $\mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathsf{T}}$ and so the predictive density is invariant under $\theta \rightarrow \theta'$, $\mathbf{C} \rightarrow \frac{\theta^2}{\theta'^2}\mathbf{C}$ (or $\mathbf{A} \rightarrow \frac{\theta}{\theta'}\mathbf{A}$).

Thus, a sensible combination of parameters to be learned would be to fix $\mathbf{C}$, eg. to be the identity, and θ, eg. to be 1, and learn the rest of the parameters. However, in the context of modelling V1 it may be more instructive to fix $\mathbf{A}$, to be a filter bank of Gabor filters, reminiscent of V1 receptive fields (though $\mathbf{A}$ really defines *projective* not receptive fields), and θ, and learn the rest of the parameters (including $\mathbf{C}$). Note that by fixing $\mathbf{A}$ we do lose some expressive power in statistical terms but may obtain more interpretable results in biological terms. Nevertheless, for completeness, we derive the learning rules both for $\mathbf{A}$ and $\mathbf{C}$. (And we don't derive the learning rule for k since there are no image data sets at our disposal that would be controlled for having natural global luminance / contrast statistics.)

We use EM (Dempster et al., 1977) to perform maximum likelihood iteratively. In each E-step we compute the (sufficient statistics of the) posterior over latent variables, $\mathbf{u}$, $\mathbf{y}$, and z, and and in each M-step we compute the values of the new parameters, ϑ^* such that they maximise the negative free energy (Neal and Hinton, 1998) given the (sufficient statistics of the) posterior computed in the previous E-step: [2]

$$\vartheta^* = \underset{\vartheta'}{\operatorname{argmax}} \sum_t \int \mathrm{d}z\, d\mathbf{u}\, \mathrm{P}(\mathbf{y}, \mathbf{u}, z | \mathbf{x}_t; \vartheta)\ \ln \mathrm{P}(\mathbf{x}_t, \mathbf{y}, \mathbf{u}, z; \vartheta') \tag{27}$$

$$= \underset{\vartheta'}{\operatorname{argmax}} \sum_t \int \mathrm{d}z\, d\mathbf{u}\ \ \mathrm{P}(\mathbf{y}, \mathbf{u}, z | \mathbf{x}_t; \vartheta) \cdot \\ \cdot \Big[\ln \mathrm{P}(\mathbf{u}; \mathbf{C}') + \ln \mathrm{P}\Big(\mathbf{x}_t | \mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) + \ldots \Big] \tag{28}$$

$$= \underset{\vartheta'}{\operatorname{argmax}}\ \ \sum_t \int \mathrm{d}\mathbf{u}\, \mathrm{P}(\mathbf{u} | \mathbf{x}_t; \vartheta)\ \ln \mathrm{P}(\mathbf{u}; \mathbf{C}') + \\ + \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y} | \mathbf{x}_t; \vartheta) \ln \mathrm{P}\Big(\mathbf{x}_t | \mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) + \ldots \tag{29}$$

4.1 M-step for C

Preliminaries:

$$\ln \mathrm{P}(\mathbf{u}; \mathbf{C}) \propto \ln \left|\mathbf{C}^{-1}\right| - \mathbf{u}^\mathsf{T} \mathbf{C}^{-1} \mathbf{u} + \ldots \tag{30}$$

$$\frac{\partial}{\partial C_{ij}^{-1}} \ln \mathrm{P}(\mathbf{u}; \mathbf{C}) \propto \mathrm{Tr}\left(\mathbf{C}\, \frac{\partial \mathbf{C}^{-1}}{C_{ij}^{-1}} \right) - u_i\, u_j = C_{ij} - u_i\, u_j \tag{31}$$

$$\nabla_{\mathbf{C}^{-1}} \ln \mathrm{P}(\mathbf{u}; \mathbf{C}) \propto \mathbf{C} - \mathbf{u}\, \mathbf{u}^\mathsf{T} \tag{32}$$

[2] Thus we are performing *complete* M steps to reach the minimum corresponding to the posterior in each iteration as opposed to incomplete M-steps which would just move us slightly towards that minimum – for a discussion of these issues see Neal and Hinton, 1998

And so:

$$\mathbf{C}^* = \operatorname*{argmax}_{\mathbf{C}'} \sum_t \int \mathrm{d}\mathbf{u}\, \mathrm{P}(\mathbf{u}|\mathbf{x}_t) \ln \mathrm{P}(\mathbf{u}; \mathbf{C}') \tag{33}$$

$$\mathbf{0} = \nabla_{\mathbf{C}'^{-1}}\Big|_{\mathbf{C}'^{-1}=\mathbf{C}^{*-1}} \sum_t \int \mathrm{d}\mathbf{u}\, \mathrm{P}(\mathbf{u}|\mathbf{x}_t) \ln \mathrm{P}(\mathbf{u}; \mathbf{C}') \tag{34}$$

$$= \sum_t \int \mathrm{d}\mathbf{u}\, \mathrm{P}(\mathbf{u}|\mathbf{x}_t) \;\; \nabla_{\mathbf{C}'^{-1}}\Big|_{\mathbf{C}'^{-1}=\mathbf{C}^{*-1}} \ln \mathrm{P}(\mathbf{u}; \mathbf{C}') \tag{35}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \int \mathrm{d}\mathbf{u}\, \mathrm{P}(\mathbf{u}|z, \mathbf{x}_t) \left(\mathbf{C}^* - \mathbf{u}\,\mathbf{u}^\mathsf{T}\right) \tag{36}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[\mathbf{C}^* - \left(\langle \mathbf{u}|z, \mathbf{x}_t \rangle \left\langle \mathbf{u}^\mathsf{T}|z, \mathbf{x}_t \right\rangle + \mathrm{Cov}[\mathbf{u}|z, \mathbf{x}_t]\right)\right] \tag{37}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[\mathbf{C}^* - \left(\boldsymbol{\mu}(z, \mathbf{x}_t)\, \boldsymbol{\mu}^\mathsf{T}(z, \mathbf{x}_t) + \boldsymbol{\Sigma}(z)\right)\right] \tag{38}$$

$$= T\, \mathbf{C}^* - \boldsymbol{\Xi}_{\mathrm{u}} \tag{39}$$

with

$$\boldsymbol{\Xi}_{\mathrm{u}} = \sum_z \boldsymbol{\lambda}(z) \tag{40}$$

$$\boldsymbol{\lambda}(z) = \boldsymbol{\Sigma}(z)\; l(z) + \sum_t \mathrm{P}(z|\mathbf{x}_t)\; \boldsymbol{\mu}(z, \mathbf{x}_t)\, \boldsymbol{\mu}^\mathsf{T}(z, \mathbf{x}_t) \tag{41}$$

$$l(z) = \sum_t \mathrm{P}(z|\mathbf{x}_t) \tag{42}$$

$$\mathbf{C}^* = \frac{1}{T}\, \boldsymbol{\Xi}_{\mathrm{u}} \tag{43}$$

4.2 M-step for A

Preliminaries:

$$\ln \mathrm{P}\big(\mathbf{x}|\mathbf{y}; \sigma_{\mathrm{x}}^2, \mathbf{A}\big) \propto -\left(\mathbf{x} - \mathbf{A}\mathbf{y}\right)^\mathsf{T} \left(\mathbf{x} - \mathbf{A}\mathbf{y}\right) + \ldots = \sum_i \left(x_i - \sum_j A_{ij} y_j\right)^2 \tag{44}$$

$$\frac{\partial}{\partial A_{ij}} \ln \mathrm{P}\big(\mathbf{x}|\mathbf{y}; \sigma_{\mathrm{x}}^2, \mathbf{A}\big) \propto -\left(x_i - \sum_{j'} A_{ij'} y_{j'}\right)\; y_j \tag{45}$$

$$\nabla_{\mathbf{A}} \ln \mathrm{P}\big(\mathbf{x}|\mathbf{y}; \sigma_{\mathrm{x}}^2, \mathbf{A}\big) \propto -\left(\mathbf{x} - \mathbf{A}\mathbf{y}\right)\; \mathbf{y}^\mathsf{T} \tag{46}$$

And so:

$$\mathbf{A}^* = \underset{\mathbf{A}'}{\mathrm{argmax}} \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{47}$$

$$\mathbf{0} = \left.\nabla_{\mathbf{A}'}\right|_{\mathbf{A}'=\mathbf{A}^*} \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{48}$$

$$= \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \ \left.\nabla_{\mathbf{A}'}\right|_{\mathbf{A}'=\mathbf{A}^*} \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{49}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t)\, (\mathbf{x}_t - \mathbf{A}^*\mathbf{y})\ \mathbf{y}^\mathsf{T} \tag{50}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[\mathbf{x}_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t)\, \mathbf{y}^\mathsf{T} - \mathbf{A}^* \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t)\, \mathbf{y}\mathbf{y}^\mathsf{T}\right] \tag{51}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[\mathbf{x}_t \left\langle \mathbf{y}^\mathsf{T}|z, \mathbf{x}_t\right\rangle - \mathbf{A}^* \left(\langle \mathbf{y}|z, \mathbf{x}_t\rangle \left\langle \mathbf{y}^\mathsf{T}|z, \mathbf{x}_t\right\rangle + \mathrm{Cov}[\mathbf{y}|z, \mathbf{x}_t]\right)\right] \tag{52}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[\mathbf{x}_t\, z\, \boldsymbol{\mu}^\mathsf{T}(z, \mathbf{x}_t) - \mathbf{A}^*\, z^2\ \left(\boldsymbol{\mu}(z, \mathbf{x}_t)\, \boldsymbol{\mu}^\mathsf{T}(z, \mathbf{x}_t) + \boldsymbol{\Sigma}(z)\right)\right] \tag{53}$$

$$= \boldsymbol{\Psi}_{\mathrm{y}} - \mathbf{A}^*\, \boldsymbol{\Xi}_{\mathrm{y}} \tag{54}$$

with

$$\boldsymbol{\Psi}_{\mathrm{y}} = \sum_t \mathbf{x}_t \sum_z \mathrm{P}(z|\mathbf{x}_t)\ z\, \boldsymbol{\mu}^\mathsf{T}(z, \mathbf{x}_t) \tag{55}$$

$$\boldsymbol{\Xi}_{\mathrm{y}} = \sum_z z^2\, \boldsymbol{\lambda}(z) \tag{56}$$

and $\boldsymbol{\lambda}(z)$ as defined in Eq. 41

$$\mathbf{A}^* = \boldsymbol{\Psi}_{\mathrm{y}}\ \boldsymbol{\Xi}_{\mathrm{y}}^{-1} \tag{57}$$

4.3 M-step for σ_{x}^2

Preliminaries:

$$\ln \mathrm{P}\Big(\mathbf{x}|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}\Big) = -\frac{1}{2\sigma_{\mathrm{x}}^2}\, (\mathbf{x} - \mathbf{A}\mathbf{y})^\mathsf{T}\, (\mathbf{x} - \mathbf{A}\mathbf{y}) - \frac{N}{2}\, \ln \sigma_{\mathrm{x}}^2 + \ldots \tag{58}$$

$$\frac{\partial}{\partial \sigma_{\mathrm{x}}^2} \ln \mathrm{P}\big(\mathbf{x}|\mathbf{y}; \sigma_{\mathrm{x}}^2, \mathbf{A}\big) \propto \frac{1}{\sigma_{\mathrm{x}}^4}\ (\mathbf{x} - \mathbf{A}\mathbf{y})^\mathsf{T}\, (\mathbf{x} - \mathbf{A}\mathbf{y}) - \frac{N}{\sigma_{\mathrm{x}}^2} \tag{59}$$

with $N = \dim(\mathbf{x})$.

And so:

$$\sigma_{\mathrm{x}}^{*2} = \underset{\sigma_{\mathrm{x}}^{2\prime}}{\operatorname{argmax}} \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{60}$$

$$\mathbf{0} = \left.\frac{\partial}{\partial \sigma_{\mathrm{x}}^{2\prime}}\right|_{\sigma_{\mathrm{x}}^{2\prime}=\sigma_{\mathrm{x}}^{*2}} \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{61}$$

$$= \sum_t \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|\mathbf{x}_t) \left.\frac{\partial}{\partial \sigma_{\mathrm{x}}^{2\prime}}\right|_{\sigma_{\mathrm{x}}^{2\prime}=\sigma_{\mathrm{x}}^{*2}} \ln \mathrm{P}\Big(\mathbf{x}_t|\mathbf{y}; \sigma_{\mathrm{x}}^{2\prime}, \mathbf{A}'\Big) \tag{62}$$

$$= \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t) \left[\frac{1}{\sigma_{\mathrm{x}}^{*4}} \left(\mathbf{x}_t - \mathbf{A}\mathbf{y}\right)^{\mathsf{T}} \left(\mathbf{x}_t - \mathbf{A}\mathbf{y}\right) - \frac{N}{\sigma_{\mathrm{x}}^{*2}}\right] \tag{63}$$

$$= \frac{T\,N}{\sigma_{\mathrm{x}}^{*2}} - \frac{1}{\sigma_{\mathrm{x}}^{*4}} \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t) \Big[\mathbf{x}_t^{\mathsf{T}}\mathbf{x}_t + (\mathbf{A}\mathbf{y})^{\mathsf{T}} (\mathbf{A}\mathbf{y}) - 2\,\mathbf{y}^{\mathsf{T}}\,\mathbf{A}^{\mathsf{T}}\,\mathbf{x}_t\Big] \tag{64}$$

$$= \frac{T\,N}{\sigma_{\mathrm{x}}^{*2}} - \frac{1}{\sigma_{\mathrm{x}}^{*4}} \left[\sum_t \mathbf{x}_t^{\mathsf{T}}\mathbf{x}_t + \sum_z \mathrm{P}(z|\mathbf{x}_t) \int \mathrm{d}\mathbf{y}\, \mathrm{P}(\mathbf{y}|z, \mathbf{x}_t) \left[\mathbf{y}^{\mathsf{T}}\,\mathbf{A}^{\mathsf{T}}\mathbf{A}\,\mathbf{y} - 2\,\mathbf{y}^{\mathsf{T}}\,\mathbf{A}^{\mathsf{T}}\,\mathbf{x}_t\right]\right] \tag{65}$$

$$= \frac{T\,N}{\sigma_{\mathrm{x}}^{*2}} - \frac{1}{\sigma_{\mathrm{x}}^{*4}}\, \sigma_{\|x\|}^2 \tag{66}$$

with

$$\sigma_{\|x\|}^2 = \sum_t \mathbf{x}_t^{\mathsf{T}}\mathbf{x}_t + \sum_z z^2\, \mathrm{Tr}\big(\mathbf{A}^{\mathsf{T}}\mathbf{A}\,\mathbf{\Sigma}(z)\big)\, l(z) +$$

$$+ \sum_t \sum_z \mathrm{P}(z|\mathbf{x}_t) \left[z^2\, \boldsymbol{\mu}^{\mathsf{T}}(z, \mathbf{x}_t)\ \mathbf{A}^{\mathsf{T}}\mathbf{A}\,\boldsymbol{\mu}(z, \mathbf{x}_t) - 2\,z\,\boldsymbol{\mu}^{\mathsf{T}}(z, \mathbf{x}_t)\ \mathbf{A}^{\mathsf{T}}\,\mathbf{x}_t\right] \tag{67}$$

and $l(z)$ as defined in Eq. 42

$$\sigma_{\mathrm{x}}^{*2} = \frac{\sigma_{\|x\|}^2}{T\,N} \tag{68}$$

References

Wainwright MJ, Simoncelli EP (2000) Scale mixtures of Gaussians and the statistics of natural images. In: *Adv. Neur. Inf. Proc. Syst. 12*. MIT Press, pp 855–861.

Bishop CM (2006) . Pattern recognition and machine learning. New York: Springer.

Dempster A, Laird N, Rubin D, et al. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1–38.

Neal RM, Hinton GE (1998) A view of the em algorithm that justifies incremental, sparse, and other variants. In: *Learning in graphical models* (Jordan MI, ed.), Dordrecht, The Netherlands: Kluwer, pp 355–368.