Supplementary material to
Expectation Propagation for Likelihoods Depending on an Inner Product of Two Multivariate Random Variables

Tomi Peltola
tomi.peltola@aalto.fi
Department of Biomedical Engineering
and Computational Science
Aalto University, Finland

Pasi Jylänki
pasi.jylanki@ru.nl
Donders Institute for Brain,
Cognition and Behaviour
Radboud University Nijmegen,
The Netherlands

Aki Vehtari
aki.vehtari@aalto.fi
Department of Biomedical Engineering
and Computational Science
Aalto University, Finland

1 DERIVATION OF THE TILTED DISTRIBUTION MOMENT INTEGRALS

Extended derivations of the main results in Section 3 in the main text with regard to the tilted distribution moment integrals are presented in the following.

1.1 Manipulation of the Tilted Distribution Moment Integrals Using an Integral Representation of the Dirac Delta Function

Here, the derivation of Equation 3 in the main text is justified by interpreting the integral representation of the Dirac delta (generalized) function as a limit of a sequence of functions and changing the integration order. We consider the Gaussian and probit likelihood cases.

1.1.1 Background

We represent the Dirac delta as a limit of a sequence of functions, following Olver et al. [1, p. 37–38]:

$$\delta_n(\xi) = \frac{1}{2\pi} \int \exp(-\frac{t^2}{4n}) \exp(\imath \xi t) dt. \tag{1}$$

Specifically, integrals involving the Dirac delta are interpreted as

$$\int \delta(x-a)\phi(x)dx = \lim_{n\to\infty} \int \delta_n(x-a)\phi(x)dx = \phi(a), \tag{2}$$

with the conditions that $\phi(x)$ is continuous when $x \in (-\infty, \infty)$ and that for each a, $\int \exp(-n(x-a)^2)\phi(x)dx$ converges absolutely for all sufficiently large values of n [1, p. 37–38].

We will also use Fubini's theorem [2, p. 170–171] and Lebesgue's dominated convergence theorem [2, p. 91]. Fubini's theorem allows us to change the order of integration in a double integral of a function, when we know that the function is integrable in one of the orders (i.e., the integral of the absolute value of the function is finite in one of the orders). Lebesgue's dominated convergence theorem states that if $|f_n(x)| \le g(x)$, where $g(x)$ is integrable, and $\lim_{n\to\infty} f_n = f$ pointwise, then $\lim_{n\to\infty} \int f_n(x)dx = \int f(x)dx$.

Finally, note that $|\exp(\imath x)| = 1$ for any real number x.

1.1.2 First Manipulation: Rewriting with the Dirac Delta

The first manipulation in Equation 3 in the main text is the following:

$$\int p(y|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\exp(s(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}))q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}) \tag{3}$$

$$=\iint p(y|f)\exp(-sf)\delta(f-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})df\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}) \tag{4}$$

$$=\int\lim_{n\to\infty}\int p(y|f)\exp(-sf)\delta_n(f-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})df\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}) \tag{5}$$

$$=\frac{1}{2\pi}\int\lim_{n\to\infty}\iint p(y|f)\exp(-sf)\exp(-\frac{t^2}{4n})\exp(\imath(f-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})t)dtdf\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}), \tag{6}$$

where we have dropped the subscripts i and j, the possible dependence of the likelihood on θ, and added the $1=\exp(s(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}))$ factor needed for the probit case (for Gaussian likelihood, we can take $s=0$). The first row is the tilt normalization integral, the second a formal representation using the Dirac delta function, the third its interpretation as a limit and the fourth using the integral transformation sequence. According to the discussion above, this manipulation is justified if $\phi(f)=p(y|f)\exp(-sf)$ satisfies the conditions given for ϕ.

For the Gaussian likelihood case with $s=0$, $\phi(f)\propto\exp(-\frac{1}{2\theta}(f-y)^2)$ is continuous and $\int|\exp(-n(f-a)^2)\phi(f)|df$ is a Gaussian integral, which converges when $n>0$ and the likelihood variance parameter θ is positive.

For the probit case, $\phi(f)=\Phi(f)\exp(-sf)$ for a positive class y, where Φ is the cumulative distribution function of the standard normal distribution and we choose $s>0$. Again, $\phi(f)$ is a continuous function and

$$\int|\exp(-n(f-a)^2)\phi(f)|df=\int\exp(-n(f-a)^2)\Phi(f)\exp(-sf)df<\int\exp(-n(f-a)^2)\exp(-sf)df, \tag{7}$$

which is a converging Gaussian integral for $n>0$.

1.1.3 Second Manipulation: Changing the Integration Order

The second manipulation in Equation 3 in the main text is the change of integration order (and, here, taking the limit). We first consider changing the integration order of t and f, then take the limit and, finally, change the integration order of t and $(\boldsymbol{w},\boldsymbol{x})$.

According to Fubini's theorem, the integration order can be changed when the integral is absolutely convergent in one of the integration orders. The absolute value of the integrand over (t,f) is

$$|p(y|f)\exp(-sf)\exp(-\frac{t^2}{4n})\exp(\imath(f-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})t)|=p(y|f)\exp(-sf)\exp(-\frac{t^2}{4n}). \tag{8}$$

The integral over t is convergent for any $n>0$. For Gaussian likelihood, the integral over f is a again a converging Gaussian integral. For probit likelihood, the integral over f converges when s is chosen as explained above.

Thus, we can first integrate over f and then over t in Equation 6. Integrating over f gives

$$\frac{1}{2\pi}\int\lim_{n\to\infty}\int L(t)\exp(-\frac{t^2}{4n})\exp(-\imath\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}t)dt\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}), \tag{9}$$

where $L(t)\propto\exp(y\imath t-\frac{1}{2}\theta t^2)$ for Gaussian likelihood and $L(t)\propto\frac{y}{s-\imath t}\exp(-\frac{1}{2}(t^2+2st\imath))$ for probit likelihood. Note that $g(t)=|L(t)|$ are integrable functions independent of n and that $\exp(-\frac{t^2}{4n})\in[0,1]$ and $|\exp(-\imath\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}t)|=1$. The absolute value of the integrand over t in Equation 9 can then be bounded by $g(t)$ and according to the Lebesgue's dominated convergence theorem, the order of taking the limit and the integration over t can be changed in the equation. Taking the limit leaves

$$\frac{1}{2\pi}\iint L(t)\exp(-\imath\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}t)dt\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x}). \tag{10}$$

We appeal again to Fubini's theorem on changing the integration order of t and $(\boldsymbol{w},\boldsymbol{x})$. To this end, note that $L(t)$ is absolutely integrable for Gaussian and probit likelihoods and that $\int\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x})$ is a converging Gaussian integral (for such selection of s that the covariance matrix is positive definite; see below). Thus, we have that

$$\int p(y|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})\exp(s(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}))q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x})=\frac{1}{2\pi}\iint L(t)\exp(-\imath\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}t)\exp(s\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})q^{\backslash}(\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x})dt, \tag{11}$$

which is the final form of Equation 3 in the main text for $s=0$, except that here $L(t,f)$ has already been integrated over f.

1.2 Gaussian Form of the Integrand over (w, x)

The integrand over $(\boldsymbol{w}, \boldsymbol{x})$ in Equation 11 can be seen to be of Gaussian form. We derive its mean, covariance and normalization constant in the following. The integrand is denoted with $C(t, \boldsymbol{w}, \boldsymbol{x})$ (when $s = 0$) in the main text. Here,

$$C(t, \boldsymbol{w}, \boldsymbol{x}) = \exp((s - \imath t)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) q^{\setminus}(\boldsymbol{w}, \boldsymbol{x}) \tag{12}$$

$$= (2\pi)^{-K}(|\boldsymbol{\Gamma}_w^{\setminus}||\boldsymbol{\Gamma}_x^{\setminus}|)^{\frac{1}{2}} \exp((s - \imath t)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) \exp(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_w^{\setminus})^{\mathrm{T}}\boldsymbol{\Gamma}_w^{\setminus}(\boldsymbol{w} - \boldsymbol{m}_w^{\setminus})) \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_x^{\setminus})^{\mathrm{T}}\boldsymbol{\Gamma}_x^{\setminus}(\boldsymbol{x} - \boldsymbol{m}_x^{\setminus})). \tag{13}$$

This is an unnormalized Gaussian form in the real variables $\boldsymbol{w}$ and $\boldsymbol{x}$ with complex mean and covariance. We refer to page 10 of Neretin [3] for the result showing that the integral over $\boldsymbol{w}$ and $\boldsymbol{x}$ behaves similarly to the common real parameter version with the condition that the real-part of the covariance matrix is positive definite.

If $p(\boldsymbol{x})$ is a Gaussian density, we can find its mean, for example, by solving $\frac{\partial}{\partial \boldsymbol{x}} \log p(\boldsymbol{x}) = \boldsymbol{0}$ for $\boldsymbol{x}$, and the covariance as $-[\frac{\partial^2}{\partial \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}} \log p(\boldsymbol{x})]^{-1}$. For the Gaussian form 13 in the concatenated variable $[\boldsymbol{w}^{\mathrm{T}}\ \boldsymbol{x}^{\mathrm{T}}]^{\mathrm{T}}$, we find:

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{w}} \log C(t, \boldsymbol{w}, \boldsymbol{x}) \\ \frac{\partial}{\partial \boldsymbol{x}} \log C(t, \boldsymbol{w}, \boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} -(\imath t - s)\boldsymbol{x} - \boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{w} + \boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{m}_w^{\setminus} \\ -(\imath t - s)\boldsymbol{w} - \boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{x} + \boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{m}_x^{\setminus} \end{bmatrix}, \tag{14}$$

which can be solved at zero to give the mean

$$\begin{bmatrix} \bar{\boldsymbol{m}}_w(t) \\ \bar{\boldsymbol{m}}_x(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{\Sigma}(t)^{\mathrm{T}}(\boldsymbol{m}_w^{\setminus} - (\imath t - s)(\boldsymbol{\Gamma}_w^{\setminus})^{-1}\boldsymbol{m}_x^{\setminus}) \\ \boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{\Sigma}(t)(\boldsymbol{m}_x^{\setminus} - (\imath t - s)(\boldsymbol{\Gamma}_x^{\setminus})^{-1}\boldsymbol{m}_w^{\setminus}) \end{bmatrix}, \tag{15}$$

where $\boldsymbol{\Sigma}(t) = (\boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{\Gamma}_x^{\setminus} - (\imath t - s)^2\boldsymbol{I})^{-1}$ and which is the same as given in Equation 4 in the main text (except that $\imath t$ is replaced with $(\imath t - s)$ when $s \neq 0$). The covariance is solved by computing $\frac{\partial^2}{\partial \boldsymbol{w}\partial \boldsymbol{w}^{\mathrm{T}}} \log C(t, \boldsymbol{w}, \boldsymbol{x})$, $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\mathrm{T}}} \log C(t, \boldsymbol{w}, \boldsymbol{x})$ and $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{w}^{\mathrm{T}}} \log C(t, \boldsymbol{w}, \boldsymbol{x})$. The resulting covariance matrix is

$$\bar{\boldsymbol{V}}(t) = \begin{bmatrix} \boldsymbol{\Gamma}_w^{\setminus} & (\imath t - s)\boldsymbol{I} \\ (\imath t - s)\boldsymbol{I} & \boldsymbol{\Gamma}_x^{\setminus} \end{bmatrix}^{-1}, \tag{16}$$

which is the same as given in Equation 5 in the main text (except that $\imath t$ has again been replaced with $(\imath t - s)$).

The normalization constant for the Gaussian form in Equation 13 is

$$D(t) = \int C(t, \boldsymbol{w}, \boldsymbol{x}) d(\boldsymbol{w}, \boldsymbol{x}) = (|\boldsymbol{\Gamma}_w^{\setminus}||\boldsymbol{\Gamma}_x^{\setminus}||\bar{\boldsymbol{V}}(t)|)^{\frac{1}{2}} \exp(-\frac{1}{2}d(t)), \tag{17}$$

where

$$d(t) = -(\imath t - s)^2((\boldsymbol{m}_w^{\setminus})^{\mathrm{T}}\boldsymbol{\Sigma}(t)\boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{m}_w^{\setminus} + (\boldsymbol{m}_x^{\setminus})^{\mathrm{T}}\boldsymbol{\Sigma}(t)^{\mathrm{T}}\boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{m}_x^{\setminus}) + 2(\imath t - s)(\boldsymbol{m}_w^{\setminus})^{\mathrm{T}}\boldsymbol{\Sigma}(t)\boldsymbol{\Gamma}_w^{\setminus}\boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{m}_x^{\setminus}. \tag{18}$$

1.3 One-dimensional Tilted Distribution Moment Integrals

The Equation 11 is the tilted distribution normalization integral. Substituting the Gaussian form $C(t, \boldsymbol{w}, \boldsymbol{x})$ given above and integrating over $\boldsymbol{w}$ and $\boldsymbol{x}$ gives

$$\hat{Z} = \frac{1}{2\pi} \int L(t) \int C(t, \boldsymbol{w}, \boldsymbol{x}) d(\boldsymbol{w}, \boldsymbol{x}) dt = \frac{1}{2\pi} \int L(t) D(t) dt. \tag{19}$$

The mean of the tilted distribution for $\boldsymbol{w}$ is given as

$$\hat{\boldsymbol{m}}_w = \frac{1}{2\pi} \frac{1}{\hat{Z}} \int L(t) \int \boldsymbol{w} C(t, \boldsymbol{w}, \boldsymbol{x}) d(\boldsymbol{w}, \boldsymbol{x}) dt \tag{20}$$

$$= \frac{1}{2\pi} \frac{1}{\hat{Z}} \int L(t) D(t) \int \boldsymbol{w} \frac{1}{D(t)} C(t, \boldsymbol{w}, \boldsymbol{x}) d(\boldsymbol{w}, \boldsymbol{x}) dt \tag{21}$$

$$= \frac{1}{2\pi} \frac{1}{\hat{Z}} \int L(t) D(t) \bar{\boldsymbol{m}}_w(t) dt. \tag{22}$$

$\hat{\boldsymbol{m}}_x$ is computed similarly. The covariance of the tilted distribution for $\boldsymbol{w}$ is given as

$$\hat{\boldsymbol{\Gamma}}_w^{-1} = \frac{1}{2\pi}\frac{1}{\hat{Z}}\int L(t)\int \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} C(t,\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x})dt - \hat{\boldsymbol{m}}_w\hat{\boldsymbol{m}}_w^{\mathrm{T}} \tag{23}$$

$$= \frac{1}{2\pi}\frac{1}{\hat{Z}}\int L(t)D(t)\int \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \frac{1}{D(t)} C(t,\boldsymbol{w},\boldsymbol{x})d(\boldsymbol{w},\boldsymbol{x})dt - \hat{\boldsymbol{m}}_w\hat{\boldsymbol{m}}_w^{\mathrm{T}} \tag{24}$$

$$= \frac{1}{2\pi}\frac{1}{\hat{Z}}\int L(t)D(t)[\bar{\boldsymbol{V}}_w(t) + \bar{\boldsymbol{m}}_w(t)\bar{\boldsymbol{m}}_w(t)^{\mathrm{T}}]dt - \hat{\boldsymbol{m}}_w\hat{\boldsymbol{m}}_w^{\mathrm{T}}, \tag{25}$$

where $\bar{\boldsymbol{V}}_w(t)$ is the part of $\bar{\boldsymbol{V}}(t)$ corresponding to covariance of $\boldsymbol{w}$. Using block matrix inversion formula, one finds that $\bar{\boldsymbol{V}}_w(t) = \boldsymbol{\Gamma}_x^{\setminus}\boldsymbol{\Sigma}(t)$. $\hat{\boldsymbol{\Gamma}}_x^{-1}$ is computed similarly.

References

[1] Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[2] Terence Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, 2011.

[3] Yu A Neretin. *Lectures on Gaussian Integral Operators and Classical Groups*. European Mathematical Society, 2011.

2 SUPPLEMENTARY RESULTS

2.1 Sparse PCA – Gaussian Likelihood

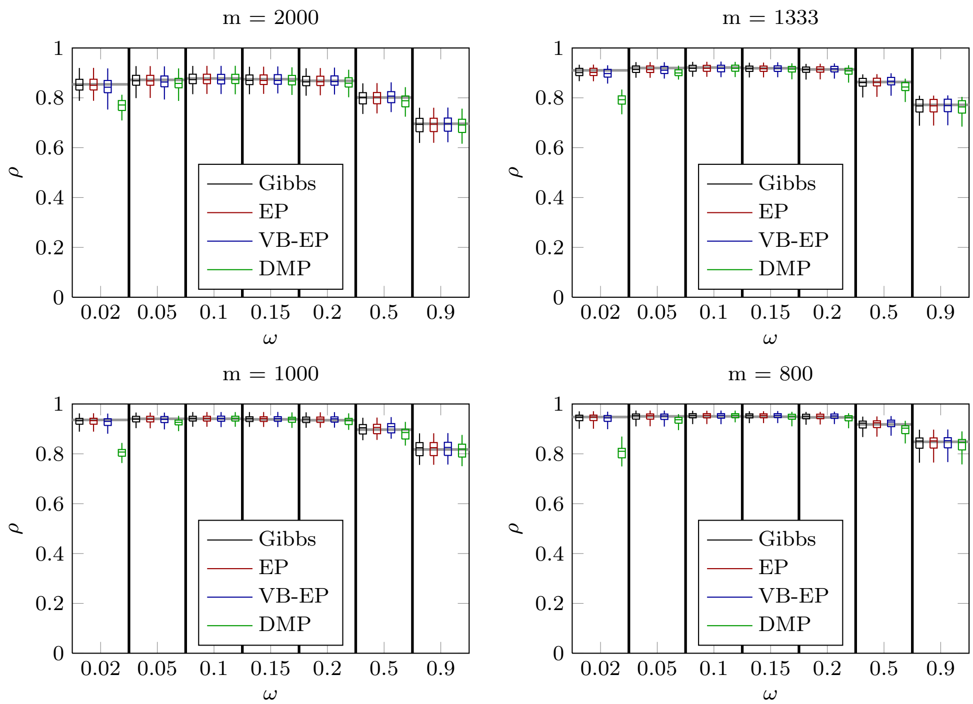

Figure 1: Boxplots of ρ over the 50 replicate datasets for Gaussian SPCA. The theoretically optimal values are shown with gray horizontal lines.

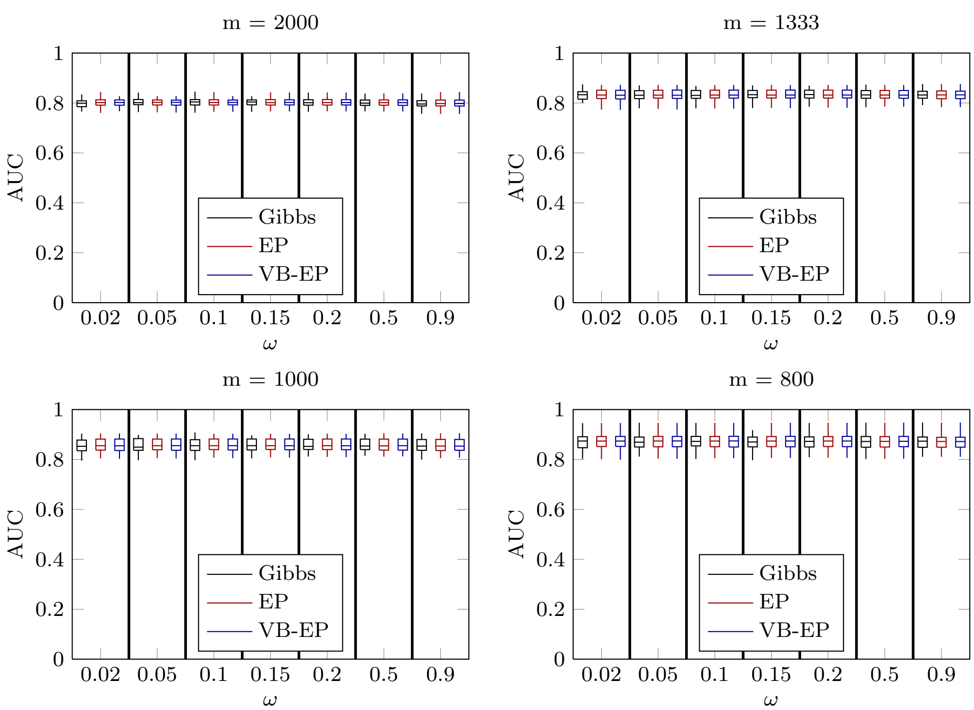

Figure 2: Boxplots of AUC over the 50 replicate datasets for Gaussian SPCA.

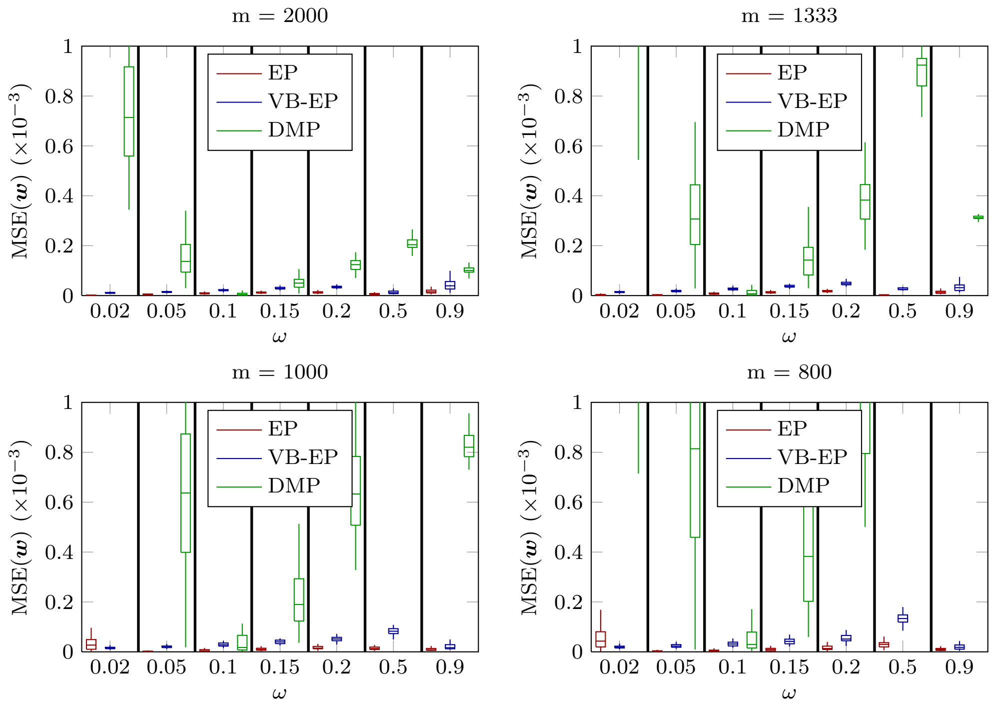

Figure 3: Boxplots of MSE($\boldsymbol{w}$) over the 50 replicate datasets for Gaussian SPCA.

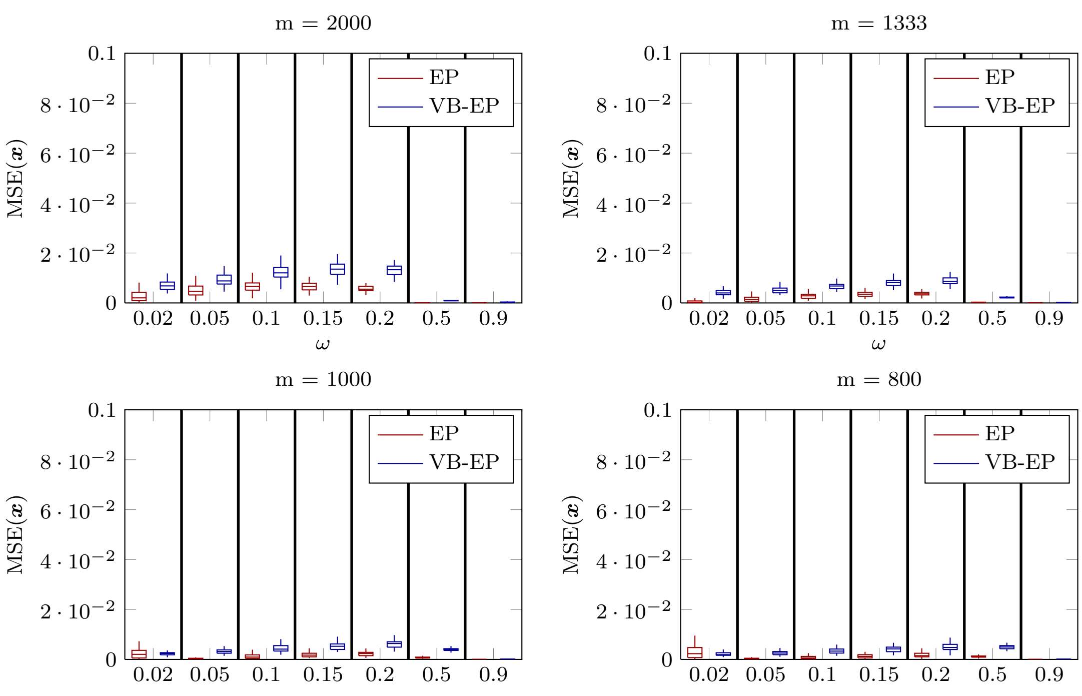

Figure 4: Boxplots of MSE($\boldsymbol{x}$) over the 50 replicate datasets for Gaussian SPCA.

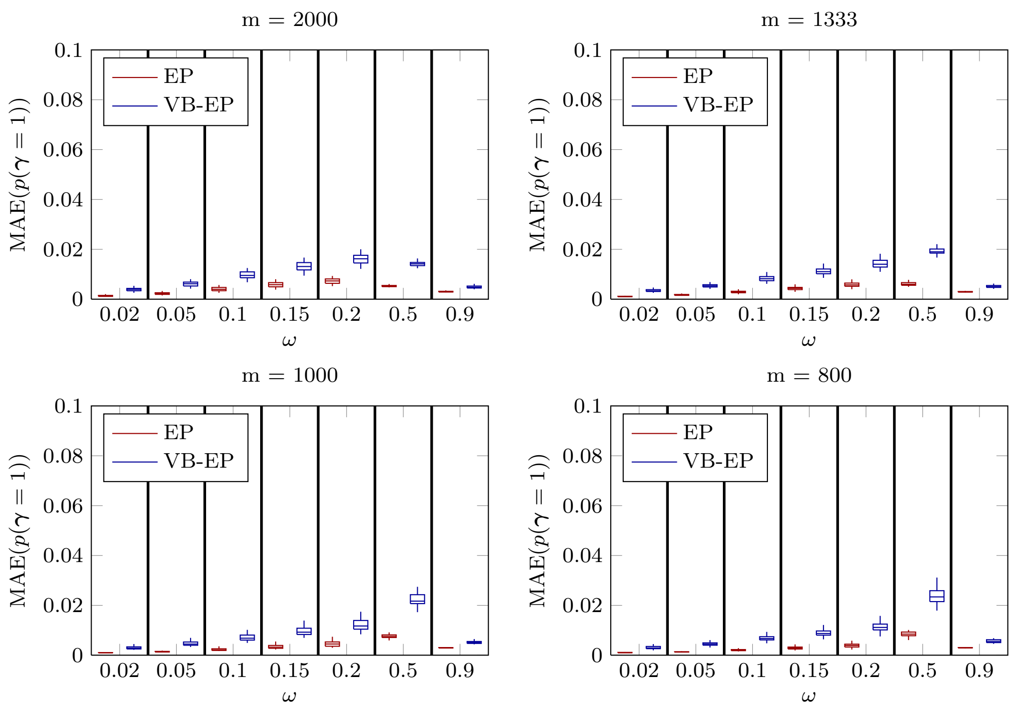

Figure 5: Boxplots of MAE($p(\boldsymbol{\gamma} = 1)$) over the 50 replicate datasets for Gaussian SPCA.

2.2 Sparse PCA – Probit Likelihood

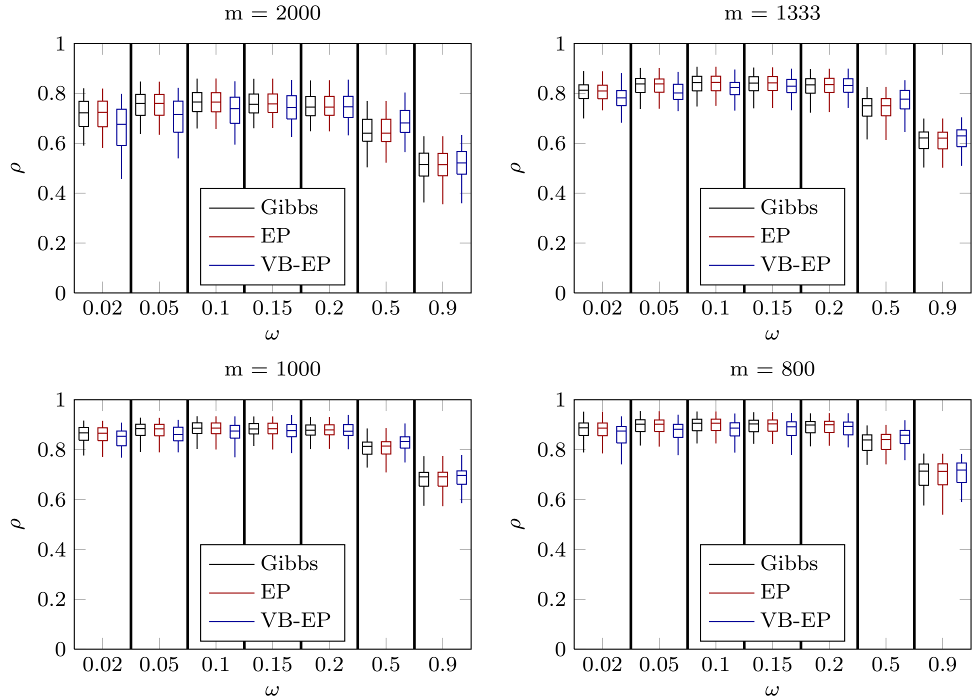

Figure 6: Boxplots of ρ over the 50 replicate datasets for Probit SPCA.

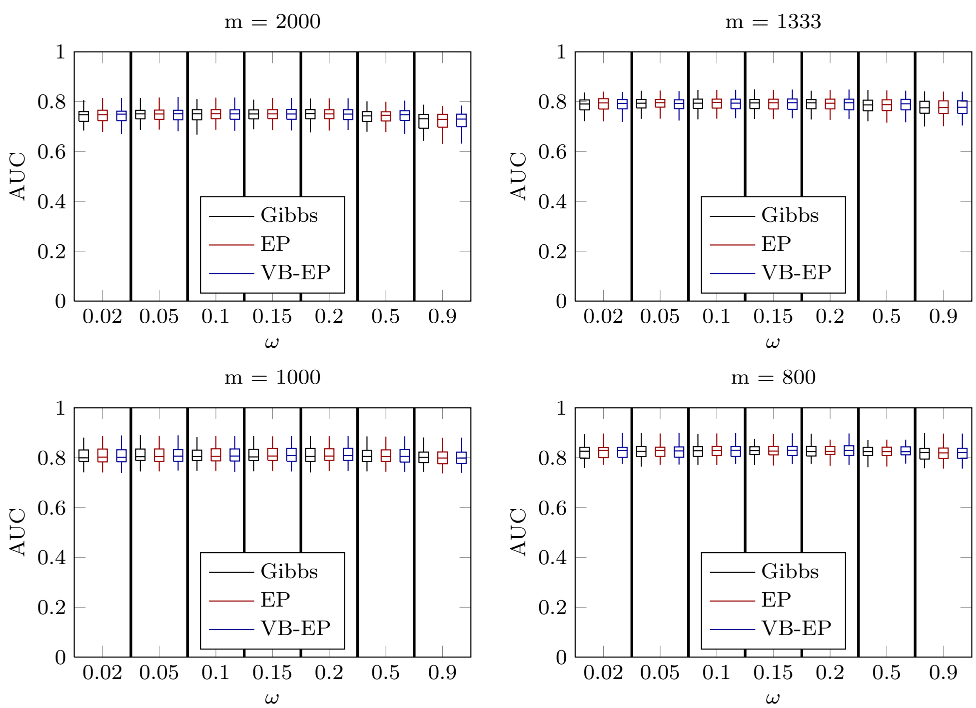

Figure 7: Boxplots of AUC over the 50 replicate datasets for Probit SPCA.

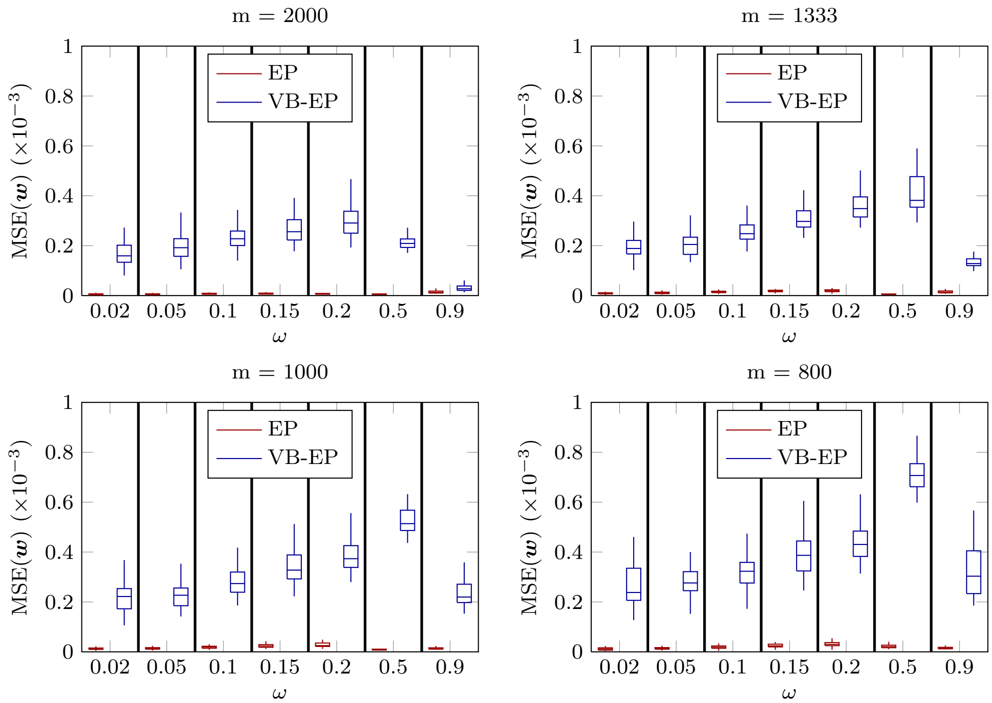

Figure 8: Boxplots of MSE($\boldsymbol{w}$) over the 50 replicate datasets for Probit SPCA.

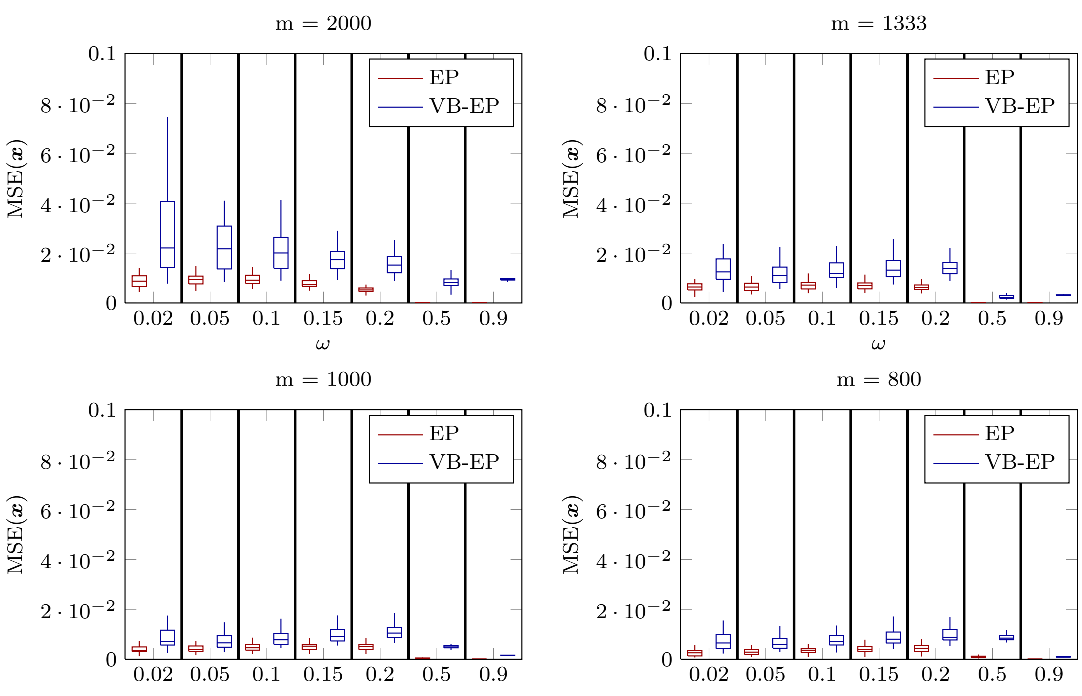

Figure 9: Boxplots of MSE($\boldsymbol{x}$) over the 50 replicate datasets for Probit SPCA.

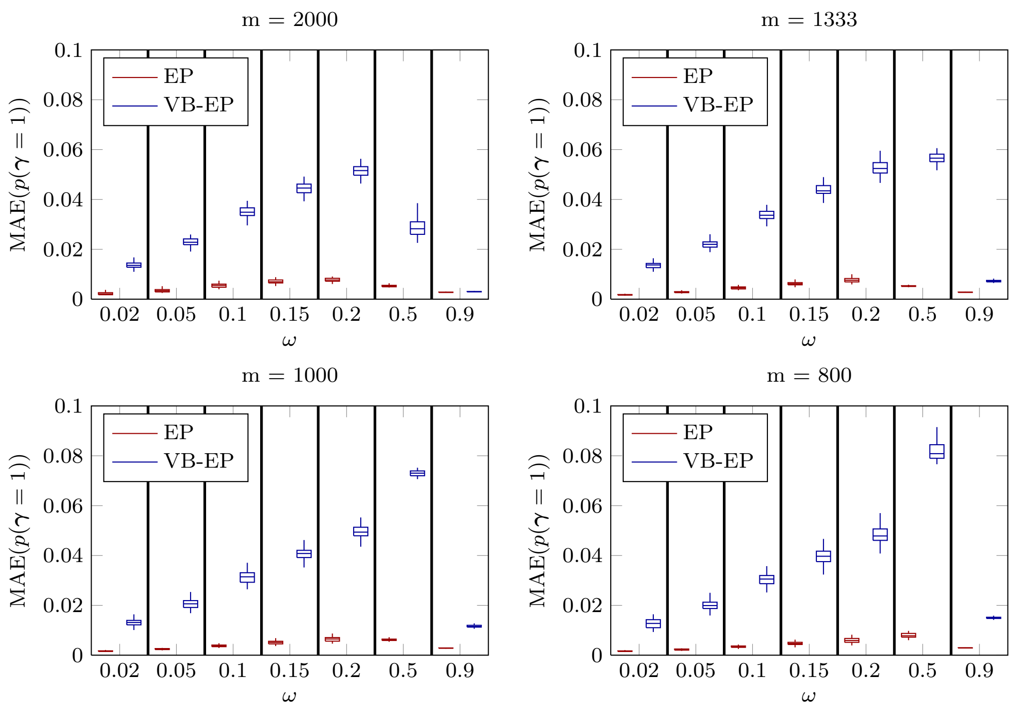

Figure 10: Boxplots of MAE($p(\boldsymbol{\gamma} = 1)$) over the 50 replicate datasets for Probit SPCA.