Some extensions of score matching

Aapo Hyvärinen

Helsinki Institute for Information Technology
and
Dept of Computer Science, University of Helsinki
Finland

Abstract

Many probabilistic models are only defined up to a normalization constant. This makes maximum likelihood estimation of the model parameters very difficult. Typically, one then has to resort to Markov Chain Monte Carlo methods, or approximations of the normalization constant. Previously, a method called score matching was proposed for computationally efficient yet (locally) consistent estimation of such models. The basic form of score matching is valid, however, only for models which define a differentiable probability density function over $\mathbb{R}^n$. Therefore, some extensions of the framework are proposed. First, a related method for binary variables is proposed. Second, it is shown how to estimate non-normalized models defined in the non-negative real domain, i.e. $\mathbb{R}^n_+$. As a further result, it is shown that the score matching estimator can be obtained in closed form for some exponential families.

Key words: Statistical estimation, non-normalized models, score matching, partition function, Markov Chain Monte Carlo

1 Introduction

In machine learning, statistics, or signal processing, one often wants to estimate statistical models which cannot be easily normalized. Let us denote the observed data vector by $\mathbf{x}$, and a parameter vector by $\boldsymbol{\theta}$. The data vector $\mathbf{x}$ can take either discrete or continuous values in a domain D. The problem we consider here is that the probability distribution function or the probability density function (both abbreviated as pdf) $p(\mathbf{x}; \boldsymbol{\theta})$ is only known up to a

* Contact address: Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358-9-191 51120, email: aapo.hyvarinen@helsinki.fi

normalization constant:

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\mathbf{x};\boldsymbol{\theta}). \tag{1}$$

That is, we know how to compute q efficiently (typically, using an analytical formula), but we do not know how to compute Z. In principle, Z is obtained by integrating over the domain D:

$$Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in D} q(\boldsymbol{\xi};\boldsymbol{\theta}) d\boldsymbol{\xi}, \tag{2}$$

where the integral is replaced by summation if D is discrete.

Maximum likelihood estimation of the parameter vector $\boldsymbol{\theta}$ is not possible without computation of the normalization constant, also called the partition function. However, computing the integral in (2) is computationally hard. Typically used methods are either computationally very complex, e.g. Markov Chain Monte Carlo methods, or they are based on approximations that may be inconsistent in the general case, e.g. pseudo-likelihood [1], contrastive divergence [2].

For the case of continuous-valued data, $D = \mathbb{R}^n$, a new approach was proposed in [3]. The new method, called score matching, completely avoids the computation of the normalization constant, but provably provides a estimator that is consistent. The idea is to consider the gradients of the log-derivatives of the densities given by the model, and by the observed data distribution. Let us denote the gradient of the log-pdf given by the model by $\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta})$:

$$\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\boldsymbol{\xi};\boldsymbol{\theta}) \\ \vdots \\ \psi_n(\boldsymbol{\xi};\boldsymbol{\theta}) \end{pmatrix} = \nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi};\boldsymbol{\theta}). \tag{3}$$

This function was called, with a slight abuse of conventional terminology, the "score function" in [3], because it is the Fisher score function with respect to a hypothetical location parameter: Assuming an additional location parameter vector $\boldsymbol{\mu}$, we obtain $\boldsymbol{\psi}$ when we take the gradient of the log-pdf $\log p(\boldsymbol{\xi} - \boldsymbol{\mu};\boldsymbol{\theta})$ with respect to $\boldsymbol{\mu}$ and evaluate it at $\boldsymbol{\mu} = \mathbf{0}$. Likewise, we denote by $\boldsymbol{\psi}_{\mathbf{x}}(.) = \nabla_{\boldsymbol{\xi}} \log p_{\mathbf{x}}(.)$ the gradient of the logarithm of the pdf $p_{\mathbf{x}}$ of the observed data $\mathbf{x}$.

The point in using this function $\boldsymbol{\psi}$ is that it does not depend on $Z(\boldsymbol{\theta})$ at all: the normalization constant disappears when taking the derivative of the logarithm with respect to $\boldsymbol{\xi}$ (i.e. the data variable).

Thus, it was proposed in [3] that the model is estimated by minimizing the

expected squared distance between the model "score function" $\boldsymbol{\psi}(.;\boldsymbol{\theta})$ and the data "score function" $\boldsymbol{\psi}_{\mathbf{x}}(.)$. We define this squared distance as

$$J_{SM}(\boldsymbol{\theta}) = \frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi})\|\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi}. \tag{4}$$

Then, the *score matching* estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J_{SM}(\boldsymbol{\theta}).$$

This approach may not seem to provide any computational advantage at first sight, because the expression in (4) contains $\boldsymbol{\psi}_{\mathbf{x}}(.)$, the gradient of the logarithm of the data pdf, which is difficult to estimate. However, it was proven in [3] that using partial integration, J can be brought to an easily computable form:

$$J_{SM}(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi}\in\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \left[\partial_i\psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) + \frac{1}{2}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})^2\right] d\boldsymbol{\xi} + \text{const.} \tag{5}$$

where the constant does not depend on $\boldsymbol{\theta}$, and

$$\partial_i\psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\xi_i} = \frac{\partial^2 \log q(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\xi_i^2} \tag{6}$$

is the second partial derivative of the model log-pdf with respect to the i-th variable. A sample version of this objective function is easily computed by replacing the integration over $p_{\mathbf{x}}$ by a sample average and ignoring the constant, which gives

$$\tilde{J}_{SM}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \partial_i\psi_i(\mathbf{x}(t);\boldsymbol{\theta}) + \frac{1}{2}\psi_i(\mathbf{x}(t);\boldsymbol{\theta})^2. \tag{7}$$

where we have a sample $\mathbf{x}(1),\ldots,\mathbf{x}(T)$ of observations, T denoting sample size.

The utility of going from (4) to (5) and finally to (7) is that (5) no longer contains any derivatives of the unknown data pdf $p_{\mathbf{x}}$ (e.g. no $\boldsymbol{\psi}_{\mathbf{x}}$), and the sample version (7) does not contain the unknown data pdf at all. Thus, the sample version in (7) can be computed as a simple sample average of functions which are typically readily calculable in closed form.

In this paper, we introduce some extensions of the basic score matching method. An important restriction of the original score matching methods is that the pdf's must be differentiable over the whole real space $D = \mathbb{R}^n$. In particular, the variables must be continuous-valued. In this paper, we lift some of these restrictions. In Section 2 we develop an analoguous method in the

case where the data takes values in the binary space, $D = \{-1, 1\}^n$. In Section 3, we show how score matching can be used, with some changes in the objective function, when the data is non-negative, i.e. has a differentiable pdf in $D = \mathbb{R}_+^n$. Furthermore, in Section 4 we analyze the original score matching method in the case of an exponential density family, and show how the estimator can then be obtained in closed form in some cases.

2 Estimation of binary models by ratio matching

2.1 Construction and analysis of estimator

In this section, we generalize score matching to binary variables. Actually, this generalization differs from score matching quite a lot, but it retains the basic purpose of providing a computationally simple and (locally) consistent estimation method for statistical models of binary variables where the normalization constant is not known.

To fix the notation, assume we observe an n-dimensional binary random vector $\mathbf{x} \in \{-1, +1\}^n$ which has a probability distribution function denoted by $P_{\mathbf{x}}(.)$. We have a parametrized density model $P(.; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an m-dimensional vector of parameters. We want to estimate the parameter $\boldsymbol{\theta}$ for $\mathbf{x}$, i.e. we want to approximate $P_{\mathbf{x}}(.)$ by $P(.; \hat{\boldsymbol{\theta}})$ for the estimated parameter value $\hat{\boldsymbol{\theta}}$. The model specification only gives P up to a multiplicative constant $Z(\boldsymbol{\theta})$:

$$P(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} Q(\boldsymbol{\xi}; \boldsymbol{\theta}). \tag{8}$$

In principle, Z is given by the sum:

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{\xi} \in \{-1,+1\}^n} Q(\boldsymbol{\xi}; \boldsymbol{\theta}) \tag{9}$$

whose computation is exponential in the dimension n. Thus, for any larger dimension n, direct numerical computation of Z is out of the question, just like in the continous case.

The method we propose here is based on minimizing the expected squared distance of the ratios of certain probabilities given by the model and the corresponding ratios in the observed data. We show two important properties for this new objective function. First, the distance can be estimated by a very simple formula involving only sample averages of a simple function of $Q(.; \boldsymbol{\theta})$ given by the model. Thus, the computations involved are essentially not more complicated than in the case where we know an analytical expression for the

normalization constant. Second, minimization of this objective function gives a (locally) consistent estimator.

In our method we consider ratios of probabilities. In particular, we consider the ratio of $P(\mathbf{x})$ and $P(\mathbf{x}_{-i})$ where $\mathbf{x}_{-i}$ denotes a vector in which the i-th element of $\mathbf{x}$ has been flipped:

$$\mathbf{x}_{-i} = (x_1, x_2, \ldots, -x_i, \ldots, x_n). \tag{10}$$

(If the binary values are given by 0 and 1, the minus operator will be replaced by a Boolean negation which transforms 0 to 1 and 1 to 0.)

The basic principle in our method is to force the ratios $P_{\mathbf{x}}(\mathbf{x})/P_{\mathbf{x}}(\mathbf{x}_{-i})$ to be as close as possible to the corresponding ratios given the model, i.e., $P(\mathbf{x};\boldsymbol{\theta})/P(\mathbf{x}_{-i};\boldsymbol{\theta})$. Thus, the method is called *(probability) ratio matching*. The obvious benefit of using ratios of probabilities is that they do not depend on the normalization constant. Obviously, we have

$$\frac{P(\mathbf{x})}{P(\mathbf{x}_{-i})} = \frac{Q(\mathbf{x})}{Q(\mathbf{x}_{-i})}, \tag{11}$$

where $P(\mathbf{x})$ stands for either $P_{\mathbf{x}}(\mathbf{x})$ or $P(\mathbf{x};\boldsymbol{\theta})$.

What remains to be done is to define a distance for the ratios, including some kind of weighting. An crucial point to note is that if we just use the ratios as such, there may often be division by zero: especially the observed probabilities $P_{\mathbf{x}}(\mathbf{x})$ are often zero for some $\mathbf{x}$. Thus, we prefer to consider the following transformation of the ratios:

$$g(u) = \frac{1}{1+u}. \tag{12}$$

Now, any probability that is zero and leads to a ratio that is infinite will simply give a value of $g(\infty) = 0$ for this transformation, and any numerical problems will be avoided. As for the weighting, we use the observed probabilities $P_{\mathbf{x}}(\mathbf{x})$ because this choice is natural and leads to algebraically simple expressions.

Thus, we propose that the model is estimated by minimizing the following objective function:

$$J_{RM}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{\boldsymbol{\xi} \in \{-1,+1\}^n} P_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \Big\{ \Big[g(P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})) - g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \Big]^2 + \Big[g(P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P_{\mathbf{x}}(\boldsymbol{\xi})) - g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \Big]^2 \Big\}, \tag{13}$$

where, for the sake of symmetry that is important for subsequent algebraic simplifications, we take the sum of two square distances with the roles of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_{-i}$ switched.

It may seem that this function is very difficult to compute, possibly requiring that the probabilities of all the possible values of $\mathbf{x}$ are computed and stored. Such a method would have exponential complexity and would not be of any use. Fortunately, this problem does not arise because we can actually compute J_{RM} quite simple as a sample average of certain functions of the observations. This is show by the following theorem, proven in Appendix A:

Theorem 1 *Assume that all the probabilities are non-zero. The objective function J_{RM} in (13) can be equivalently expressed as*

$$J_{RM}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\xi} \in \{-1,+1\}^n} P_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} g^2(Q(\boldsymbol{\xi}_{-i}; \boldsymbol{\theta})/Q(\boldsymbol{\xi}; \boldsymbol{\theta})) + const. \tag{14}$$

where the constant does not depend on $\boldsymbol{\theta}$, and g is as defined in (12).

Although the proof assumes that all the probabilities are non-zero, the objective function is well-behaving even in the limit of zero probabilities. Thus, for any practical purposes, the constraint of non-zero probabilities can be ignored.

Note that the Theorem cannot be obtained for any arbitrary function g. In fact, the function g has been carefully chosen in order to obtain the simplified form given by the Theorem.

Given a sample $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ of observations, where T denotes sample size, we thus propose that the model be estimated by

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \tilde{J}_{RM}(\boldsymbol{\theta}), \tag{15}$$

where $\tilde{J}_{RM}$ is the sample version of the equivalent form of J given by the theorem:

$$\tilde{J}_{RM}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} g^2(Q(\mathbf{x}_{-i}(t); \boldsymbol{\theta})/Q(\mathbf{x}(t); \boldsymbol{\theta})). \tag{16}$$

Now we see why the formula in Theorem 1 is useful. The sample version $\tilde{J}_{RM}$ can be computed as a sample average of a simple nonlinear function (square of g) of a ratio of the non-normalized probabilities Q, assumed to be easy to compute. This is stark contrast to computation of a sample version of (13), which requires computation and memory storage of the sample probabilities $P_{\mathbf{x}}(\boldsymbol{\xi})$ for all $\boldsymbol{\xi}$. Such computation and storage has, in general, exponential complexity, and therefore impossible for high-dimensional data.

As for the consistency, we have a result which is completely analoguous to the consistency theorem in [3]. This is given by the following Theorem, proven in Appendix B:

Theorem 2 *Assume that the data follows the model for some parameter value*

$\boldsymbol{\theta}^*$, i.e. $P_{\mathbf{x}}(\boldsymbol{\xi}) = P(\boldsymbol{\xi}; \boldsymbol{\theta}^*)$ for all $\boldsymbol{\xi}$. Assume that $P(\boldsymbol{\xi}; \boldsymbol{\theta}^*) > 0$ for all $\boldsymbol{\xi}$. Assume further that the model is identifiable in the sense that there is no other parameter value that gives the same distribution $P_{\mathbf{x}}$.

Then, $J_{RM}(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Furthermore, the estimator obtained by minimization of $\tilde{J}_{RM}$ is consistent, i.e. it converges in probability towards the true value of $\boldsymbol{\theta}$ when sample size approaches infinity. [1]

2.2 Example: fully visible Boltzmann machine

2.2.1 Derivation of objective function and gradient

As an example of ratio matching, we shall consider estimation of the following model

$$Q(\mathbf{x}; [\mathbf{M}, \mathbf{b}]) = \exp(\frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{b}^T\mathbf{x}), \tag{17}$$

where the parameter matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$ is symmetric with zero diagonal, the additional parameter vector $\mathbf{b}$ is n-dimensional, and $\mathbf{x}$ is binary. This is a special case ("fully visible", i.e. no latent variables) of the Boltzmann machine framework [4].

We obtain by straightforward calculation

$$\begin{aligned} &Q(\boldsymbol{\xi}_{-i}; [\mathbf{M}, \mathbf{b}])/Q(\boldsymbol{\xi}; [\mathbf{M}, \mathbf{b}]) \\ &\quad = \exp(\frac{1}{2}\boldsymbol{\xi}^T\mathbf{M}\boldsymbol{\xi} - 2\xi_i\mathbf{m}_i^T\boldsymbol{\xi} + \mathbf{b}^T\boldsymbol{\xi} - 2b_i\xi_i)/\exp(\frac{1}{2}\boldsymbol{\xi}^T\mathbf{M}\boldsymbol{\xi} + \mathbf{b}^T\boldsymbol{\xi}) \\ &\quad = \exp(-2\xi_i\mathbf{m}_i^T\boldsymbol{\xi} - 2b_i\xi_i). \end{aligned} \tag{18}$$

For notational simplicity, we shall replace the sum over the sample by the sample average operator $\hat{E}$ and drop the sample index t if there is no possibility of confusion. Thus, the objective function $\tilde{J}_{RM}$ in (16) is equal to

$$\tilde{J}_{RM}(\mathbf{M}, \mathbf{b}) = \sum_{i=1}^{n} \hat{E}\{g^2(\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i))\}. \tag{19}$$

The gradient of $\tilde{J}$ with respect to one element of $\mathbf{M}$ can be easily calculated. In Appendix C we show that it can be given in the form

$$8\nabla_{m_{ij}}\tilde{J} = \hat{E}(1 - \tanh^2(\mathbf{m}_i^T\mathbf{x} + b_i))[x_ix_j - x_j\tanh(\mathbf{m}_i^T\mathbf{x} + b_i)], \tag{20}$$

[1] From a computational viewpoint, we also have to assume that the numerical optimization algorithm used is able to find the global minimum of J_{RM}; there is no guarantee of the convexity of the objective function. This is why the consistency was called "local" in [3], and we have used the same qualification in parts of this article as well. However, if we assume that numerical optimization is perfect, there is no need to use such qualification.

which has the benefit that it shows the close connection to pseudo-likelihood derived in [5]. In fact, this gradient turns out to be a weighted version of the gradient of the pseudo-likelihood, the weighting given by $1-\tanh^2(\mathbf{m}_i^T\mathbf{x}+b_i)$.

Since $\mathbf{M}$ is constrained to be symmetric and have zero diagonal, the gradient has to be projected on this linear space. Thus, we compute

$$\tilde{\nabla}_{m_{ij}}\tilde{J} = \frac{1}{2}(\nabla_{m_{ij}}\tilde{J} + \nabla_{m_{ji}}\tilde{J}), \tag{21}$$

and update $\hat{\mathbf{M}}$ using this projected gradient in a gradient descent step:

$$\Delta\hat{m}_{ij} = -\mu\tilde{\nabla}_{m_{ij}}\tilde{J}(\hat{\mathbf{M}}), \tag{22}$$

where μ is a step size. Similarly, we can compute the derivative with respect to $\mathbf{b}$, which equals

$$\nabla_{b_i}\tilde{J} = \hat{E}(1-\tanh^2(\mathbf{m}_i^T\mathbf{x}+b_i))[x_i - \tanh(\mathbf{m}_i^T\mathbf{x}+b_i)]. \tag{23}$$

Regarding computational complexity, Equation (20) shows that the complexity of computing the gradient of the ratio matching objective function is essentially of the same order as for pseudo-likelihood. Computation of the term in brackets is common to both methods, whereas ratio matching requires the additional computation of the weights which multiply the bracketed term. The weights contain partly the same terms as the bracketed term, so the increase is computational load is typically less than doubled.

2.2.2 Simulations

We performed simulation to validate the different estimation methods for the fully visible Boltmann machine. We created random matrices $\mathbf{M}$ so that the elements had normal distributions with zero mean and variance of .25. The dimension n was set to 6 which is small enough to enable exact sampling from the distribution. The bias elements b_i were set to zero. We generated data from the distribution and estimated the parameters using ratio matching, maximum pseudo-likelihood [1,5], and maximum likelihood obtained by exact computation of the normalization constant Z. (Pseudo-likelihood is also asymptotically equivalent to contrastive divergence as shown in [5].)

We estimated the parameters for various sample sizes: 500, 1000, 2000, 4000, 8000, 16000, and 32000. For each sample size, we created 10 different data sets and ran the estimation once on each data set using a single random initial point. For each estimation, the estimation error was computed as the Euclidean distance of the real matrix $\mathbf{M}$ and its estimate. We took the mean of the logarithm of the 10 estimation errors.

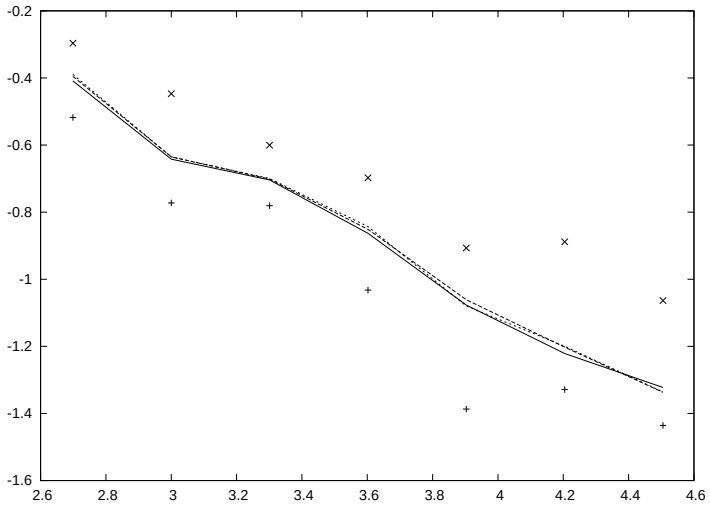

Fig. 1. The median estimation errors of ratio matching (solid line) compared with errors of maximum pseudo-likelihood/contrastive divergence estimation (dashed line) and maximum likelihood (dash-dotted line) for the fully visible Boltzmann machine. Horizontal axis: $\log_{10}$ of sample size. Vertical axis: $\log_{10}$ of estimation error. Individual crosses show the maximum and minimum errors for ratio matching among the 10 runs.

The results are shown in Figure 1. Two things are clearly seen here. First, all three estimators give extremely similar errors. This is comprehensible because the ratio matching gradient is basically the same as the pseudo-likelihood or expected contrastive divergence gradient, up to a re-weighting. We computed the weights and saw that indeed, the weights are not very different from each other, the largest being only some 20% larger than the smallest ones. The second thing we can see in Figure 1 is that ratio matching does seem to provide a consistent estimator as the error seems to converge to zero. Furthermore, optimization using a single initial point seems to pose no problems since even the maximum of errors over different runs seems to go to zero; the consistency seems to be global for this model.

3 Score matching for non-negative data

Basic score matching assumes that pdf's are differentiable over all $\mathbb{R}^n$. In this section, we show how the method can be generalized to the important case where the data are all non-negative. In other words, the pdf's are only defined in $\mathbb{R}^n_+$, or $\{\mathbb{R}_+ \cup \{0\}\}^n$. The reason why ordinary score matching cannot often be used in this case is that the pdf of non-negative data may exhibit a strong discontinuity in points where one of the variables is zero.

Basic score matching can be conceived as using a hypothetical a location parameter, say $\boldsymbol{\mu}$. The gradient of the log-pdf is then taken with respect to

this parameter, which is subsequently set to $\mathbf{0}$:

$$\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\mu}} \log p(\boldsymbol{\xi} + \boldsymbol{\mu}; \boldsymbol{\theta})|_{\boldsymbol{\mu}=\mathbf{0}} \,. \tag{24}$$

Here, we introduce a scale parameter $\boldsymbol{\sigma} \in \mathbb{R}^n_+$ instead of the location parameter, which is more natural when all the variables are non-negative. Let us denote by $\otimes$ the element-wise multiplication of two vectors. We use the following function, which we call the "scaling score function"

$$\nabla_{\boldsymbol{\sigma}} \log p(\boldsymbol{\xi} \otimes \boldsymbol{\sigma}; \boldsymbol{\theta})|_{\boldsymbol{\sigma}=\mathbf{1}} = (\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta})) \otimes \boldsymbol{\xi} = \boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) \otimes \boldsymbol{\xi}. \tag{25}$$

That is, we take the gradient of the log-pdf with respect to the scale parameter and evaluate it at the "default" value where $\sigma_i = 1$ for all i. Likewise, we can define the scaling score function of the data distribution $p_{\mathbf{x}}$. Now, we consider the squared distance between the scaling score functions of the data and the model:

$$J_{NN}(\boldsymbol{\theta}) = \frac{1}{2} \int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi}) \otimes \boldsymbol{\xi} - \boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) \otimes \boldsymbol{\xi}\|^2 d\boldsymbol{\xi}. \tag{26}$$

Just like in basic score matching, we can use a simple trick of partial integration to obtain a computationally simple expression to optimize. This is stated in the following theorem, proven in Appendix D:

Theorem 3 *Assume all the pdf's are differentiable in $\mathbb{R}^n_+$, as well as some weak regularity conditions.*[2] *Then the function in (26) can be expressed as*

$$J_{NN}(\boldsymbol{\theta}) = \int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \left[2\xi_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) + \xi_i^2 \partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})^2 \xi_i^2 \right] d\boldsymbol{\xi} + const. \tag{27}$$

where the ψ_i and $\partial_i \psi_i$ are the ordinary "score functions" (Fisher scores with respect to a hypothetical location parameter) and their derivatives as defined in (3) and (6).

The sample version of the objective function can be computed as:

$$\tilde{J}_{NN}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} 2x_i(t)\psi_i(\mathbf{x}(t); \boldsymbol{\theta}) + x_i(t)^2 \partial_i \psi_i(\mathbf{x}(t); \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{x}(t); \boldsymbol{\theta})^2 x_i(t)^2. \tag{28}$$

This sample version is easy to compute: it only contains some gradients of the (non-normalized) pdf's, so it is easy to compute. Thus, we have shown how the score matching framework can be extended to the case of data constrained non-negative.

[2] The regularity conditions are: the data pdf $p_{\mathbf{x}}(\boldsymbol{\xi})$ is differentiable in $\mathbb{R}^n_+$, the expectations $E_{\mathbf{x}}\{\|\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta})\|^2 \|\mathbf{x}\|^2\}$ and $E_{\mathbf{x}}\{\|\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})\|^2 \|\mathbf{x}\|^2\}$ are finite for any $\boldsymbol{\theta}$, and $p_{\mathbf{x}}(\boldsymbol{\xi})\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta})\xi_i^2$ goes to zero for any i and $\boldsymbol{\theta}$ when $\|\boldsymbol{\xi}\| \to \infty$ or $\|\boldsymbol{\xi}\| \to 0$

The (local) consistency of the estimator can be proven by a trivial modification of the consistency theorem of original score matching. In fact, we have a practically identical theorem of consistency that we give next.

Theorem 4 *Assume the pdf of* $\mathbf{x}$ *follows the model:* $p_{\mathbf{x}}(.) = p(.;\boldsymbol{\theta}^*)$ *for some* $\boldsymbol{\theta}^*$*. Assume further that no other parameter value gives a pdf that is equal*[3] *to* $p(.;\boldsymbol{\theta}^*)$*, and that* $q(\boldsymbol{\xi};\boldsymbol{\theta}^*) > 0$ *for all* $\boldsymbol{\xi} \in \mathbb{R}^n_+$*. Then*

$$J_{NN}(\boldsymbol{\theta}) = 0 \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*.$$

For a proof, see Appendix E. A consistency corollary similar to the one for original score matching, or in the Theorem 2 above, concludes the consistency analysis, but we omit it here for brevity.

An alternative approach would be to make a transformation of variables $y_i = \log x_i$ and use score matching in the transformed space. In fact, simple calculations show that this lead to exactly the same objective function as the approach just described.

A simulation example using this objective function will be provided in Section 4.2.

4 Closed-form solution in exponential family

4.1 Derivation of solution

Our last result is to show that for some exponential families, the score matching estimator can be obtained in closed form. In an exponential family, the pdf can be expressed in the form

$$\log p(\boldsymbol{\xi};\boldsymbol{\theta}) = \sum_{k=1}^{m} \theta_k F_k(\boldsymbol{\xi}) - \log Z(\boldsymbol{\theta}), \tag{29}$$

where Z is the normalization constant needed to make p integrate to unity, but as mentioned above, it does not need to be computed in our framework.

Here, we assume that the parameter space is $\mathbb{R}^m$, i.e. $\boldsymbol{\theta}$ can take all possible real values. In that case, we can find a direct expression for the estimator.

[3] In this theorem and its proof, equalities of pdf's are to be taken in the sense of equal almost everywhere in $\mathbb{R}^n_+$ with respect to the Lebesgue measure.

Let us denote the matrix of partial derivatives of F, i.e. its Jacobian, by $\mathbf{K}(\xi)$, with elements defined as:

$$K_{ki}(\boldsymbol{\xi}) = \frac{\partial F_k}{\partial \xi_i}, \tag{30}$$

and the needed matrix of second derivatives by

$$H_{ki}(\boldsymbol{\xi}) = \frac{\partial^2 F_k}{\partial \xi_i^2}. \tag{31}$$

Now, we have

$$\psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) = \sum_{k=1}^{m} \theta_k K_{ki}(\boldsymbol{\xi}), \tag{32}$$

and the objective function $\tilde{J}_{SM}$ in (7) becomes

$$\begin{aligned} \tilde{J}_{SM}(\boldsymbol{\theta}) &= \frac{1}{T} \sum_{t=1}^{T} \sum_i \left[\frac{1}{2} (\sum_{k=1}^{m} \theta_k K_{ki}(\mathbf{x}(t)))^2 + \sum_{k=1}^{m} \theta_k H_{ki}(\mathbf{x}(t)) \right] \\ &= \frac{1}{2} \boldsymbol{\theta}^T (\frac{1}{T} \sum_{t=1}^{T} \mathbf{K}(\mathbf{x}(t)) \mathbf{K}(\mathbf{x}(t))^T) \boldsymbol{\theta} + \boldsymbol{\theta}^T (\frac{1}{T} \sum_{t=1}^{T} \sum_i H_{ki}(\mathbf{x}(t))). \end{aligned} \tag{33}$$

This is a simple quadratic form of $\boldsymbol{\theta}$. Thus, the minimizing $\boldsymbol{\theta}$ can be easily solved by computing the gradient and setting it to zero. This gives $\hat{\boldsymbol{\theta}}$ is closed form as

$$\hat{\boldsymbol{\theta}} = - \left[\hat{E}\{\mathbf{K}(\mathbf{x}) \mathbf{K}(\mathbf{x})^T\} \right]^{-1} (\sum_i \hat{E}\{\mathbf{h}_i(\mathbf{x})\}), \tag{34}$$

where $\hat{E}$ denotes the sample average (i.e. expectation over the sample distribution), and the vector $\mathbf{h}_i(\mathbf{x})$ is the i-th column of the matrix $\mathbf{H}$ defined in (31).

A similar development is possible in the case of non-negative data. We shall not give the general equations which are obtained in the same way. Rather, we will show in the following example how the results are valid for non-negative data as well.

4.2 Example: non-negative gaussian family

As an example that illustrates both the closed-form solution for some exponential families and the version for non-negative data, let use consider the following distribution:

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = -\frac{1}{2} \mathbf{x}^T \mathbf{M} \mathbf{x} - \log Z(\mathbf{M}), \tag{35}$$

where $\mathbf{x}$ is constrained non-negative. This is a restriction of the multivariate gaussian distribution (with mean $\mathbf{0}$) to $\mathbb{R}^n_+$. The matrix $\mathbf{M}$ has to be constrained symmetric and positive-definite.

The first and second partial derivatives of the log-pdf's are equal to

$$\psi_i(\boldsymbol{\xi}; \mathbf{M}) = -\mathbf{m}_i^T \boldsymbol{\xi} \quad \text{and} \quad \partial_i \psi_i(\boldsymbol{\xi}; \mathbf{M}) = -m_{ii}, \tag{36}$$

where $\mathbf{m}_i$ is the i-th row of $\mathbf{M}$. The objective function for non-negative data in (28) is then equal to

$$\tilde{J}_{NN}(\mathbf{M}) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} 2x_i(t) \mathbf{m}_i^T \mathbf{x}(t) - x_i(t)^2 m_{ii} + \frac{1}{2} (\mathbf{m}_i^T \mathbf{x}(t))^2 x_i^2, \tag{37}$$

from which we obtain by simple algebraic manipulations

$$\tilde{J}_{NN}(\mathbf{M}) = -\sum_{i=1}^{n} 2\hat{E}\{x_i \mathbf{x}^T\} \mathbf{m}_i - \hat{E}\{x_i^2\} m_{ii} + \frac{1}{2} \mathbf{m}_i^T \hat{E}\{\mathbf{x}\mathbf{x}^T x_i^2\} \mathbf{m}_i, \tag{38}$$

from which we can solve the maximizing $\mathbf{m}_i$ as

$$\hat{\mathbf{m}}_i = [\hat{E}\{\mathbf{x}\mathbf{x}^T x_i^2\}]^{-1} \left[2\hat{E}\{x_i \mathbf{x}^T\} + \hat{E}\{x_i^2\} \mathbf{e}_i \right], \tag{39}$$

where $\mathbf{e}_i$ denotes the i-th canonical basis vector, i.e. a vector in which the i-th element is one and all others zero. This derivation considered each $\mathbf{m}_i$ separately, ignoring the constraints that $\mathbf{M}$ must be symmetric and positive definite. A slightly better estimator is probably obtained by projecting $\mathbf{M}$ on the space of symmetric positive-definite matrices. However, this does not seem necessary to demonstrate the consistency of our method, so we ignore this projection here.

We simulated data from this distribution in four dimensions. Such simulation is simply achieved (in low dimensions) by sampling from an ordinary gaussian distribution with covariance $\mathbf{M}^{-1}$, and rejecting any sample with negative values. We took 10 random matrices $\mathbf{A}$ that were created by taking uniformly distributed random variables independently for each element. The matrix $\mathbf{M}$ was then obtained as $\mathbf{A}\mathbf{A}^T$, which ensured that it is positive-definite. Badly conditioned $\mathbf{M}$'s were rejected because their estimation was too difficult.

The resulting estimation errors for increasing sample size are depicted in Fig. 2. The median estimation error seems to go to zero, which confirms the consistency of the estimator.

5 Conclusion

We extended the score matching framework [3] in three ways. First, we showed how a related method can be developed for binary data, still providing a consistent estimator. Second, we showed how to apply estimate non-normalized

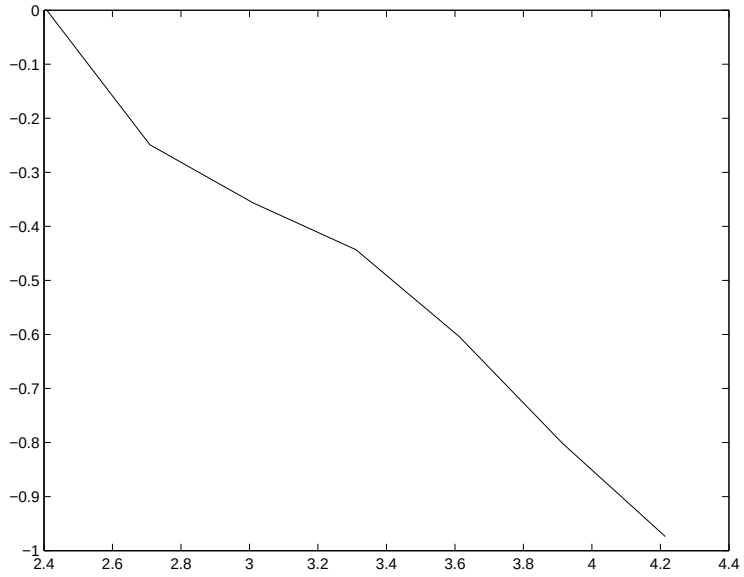

Fig. 2. The estimation errors of score matching estimation for the non-negative gaussian family. Median error was taken over ten runs. Horizontal axis: $\log_{10}$ of sample size. Vertical axis: $\log_{10}$ of estimation error.

models that are constrained to the non-negative domain. Third, we showed how the estimator can be obtained in closed form for exponential families. Related extensions were proposed in [6].

Acknowledgements

This work was supported by the Academy of Finland, Academy Research Fellow position and project #106473. I'm grateful to Shinto Eguchi and Shiro Ikeda for interesting discussions, and to Philip Dawid and Steffen Lauritzen for sharing unpublished results.

A Proof of Theorem 1

For simplicity, we do not denote the summation limits for i and $\boldsymbol{\xi}$ in the following. Note that obviously $(\boldsymbol{\xi}_{-i})_{-i} = \boldsymbol{\xi}$, and that $\boldsymbol{\xi}_{-i}$ goes though all the same values of $\boldsymbol{\xi}$, so in the sums below, we can exchange $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_{-i}$.

Taking the definition of J in (13) and doing successive algebraic manipulations

yields

$$
\begin{aligned}
2J_{RM}(\boldsymbol{\theta}) &= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[g(P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})) - g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \right]^2 \\
&+ \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[g(P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P_{\mathbf{x}}(\boldsymbol{\xi})) - g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \right]^2 \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[g(P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P_{\mathbf{x}}(\boldsymbol{\xi}))^2 + g(P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i}))^2 \right] \\
&+ \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2 + g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta}))^2 \right] \\
&- 2 \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \Big[g(P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})) g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&\qquad + g(P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P_{\mathbf{x}}(\boldsymbol{\xi})) g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \Big] . \qquad \text{(A.1)}
\end{aligned}
$$

The first term in brackets on the right-hand-side is a constant that does not depend on $\boldsymbol{\theta}$, it will be denoted by "const." below. Let us next consider the third term. It can be manipulated as follows:

$$
\begin{aligned}
&\sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \Big[g(P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})) g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&\qquad + g(P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P_{\mathbf{x}}(\boldsymbol{\xi})) g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \Big] \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \frac{P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})}{P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})} g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&+ \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \frac{P_{\mathbf{x}}(\boldsymbol{\xi})}{P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})} g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \frac{P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})}{P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})} g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&+ \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}_{-i}) \frac{P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})}{P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})} g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&= \sum_{i,\boldsymbol{\xi}} [P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})] \frac{P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})}{P_{\mathbf{x}}(\boldsymbol{\xi}) + P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})} g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}_{-i}) g(P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) = \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) g(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})/P(\boldsymbol{\xi};\boldsymbol{\theta})) \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \frac{P(\boldsymbol{\xi};\boldsymbol{\theta})}{P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) + P(\boldsymbol{\xi};\boldsymbol{\theta})} . \qquad \text{(A.2)}
\end{aligned}
$$

The second sum term on the right-hand side of (A.1) is equal to:

$$
\sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} + \frac{P(\boldsymbol{\xi};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} \right] . \qquad \text{(A.3)}
$$

So, gathering together (A.2) and (A.3), we have

$$
\begin{aligned}
2J_{RM}(\boldsymbol{\theta}) &= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} + \frac{P(\boldsymbol{\xi};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} \right] \\
&\qquad - 2\sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \frac{P(\boldsymbol{\xi};\boldsymbol{\theta})}{P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) + P(\boldsymbol{\xi};\boldsymbol{\theta})} + \text{const.} \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{P(\boldsymbol{\xi};\boldsymbol{\theta})^2 + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2 - 2P(\boldsymbol{\xi};\boldsymbol{\theta})(P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) + P(\boldsymbol{\xi};\boldsymbol{\theta}))}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} \right] + \text{const.} \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{-P(\boldsymbol{\xi};\boldsymbol{\theta})^2 - 2P(\boldsymbol{\xi};\boldsymbol{\theta})P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) - P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2 + 2P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} \right] + \text{const.} \\
&= \sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) \left[-1 + 2\frac{P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})^2}{(P(\boldsymbol{\xi};\boldsymbol{\theta}) + P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}))^2} \right] + \text{const.} \\
&= -1 + 2\sum_{i,\boldsymbol{\xi}} P_{\mathbf{x}}(\boldsymbol{\xi}) g^2(Q(\boldsymbol{\xi};\boldsymbol{\theta})/Q(\boldsymbol{\xi}_{-i};\boldsymbol{\theta})) + \text{const.}, \quad \text{(A.4)}
\end{aligned}
$$

where the last equality is due to definition of g in (12), and the interchangability of Q and P ratios as in (11). Thus, we have proven the theorem.

B Proof of Theorem 2

The hypothesis $J_{RM} = 0$, together with the assumption that $P(\boldsymbol{\xi};\boldsymbol{\theta}^*) = P_{\mathbf{x}}(\boldsymbol{\xi}) > 0$ for any $\boldsymbol{\xi}$, implies that all the ratios must be equal for the model and the observed data. This means that

$$
\begin{aligned}
P_{\mathbf{x}}(\boldsymbol{\xi})/P_{\mathbf{x}}(\boldsymbol{\xi}_{-i}) &= P(\boldsymbol{\xi};\boldsymbol{\theta})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) && \text{(B.1)} \\
&\Leftrightarrow && \text{(B.2)} \\
P_{\mathbf{x}}(\boldsymbol{\xi})/P(\boldsymbol{\xi};\boldsymbol{\theta}) &= P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}), && \text{(B.3)}
\end{aligned}
$$

for all $\boldsymbol{\xi}$ and i. Applying this identity on $\boldsymbol{\xi}_{-i}$, we have

$$
P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) = P_{\mathbf{x}}(\boldsymbol{\xi}_{-i,-k})/P(\boldsymbol{\xi}_{-i,-k};\boldsymbol{\theta}) \qquad \text{(B.4)}
$$

for some other index k. We can apply this recursively n times for any sequence of indices.

Now, fix any point $\boldsymbol{\xi}^0$. Take the set of indices for which $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^0$ differ, and use this recursion on that set of indices. We get

$$
P_{\mathbf{x}}(\boldsymbol{\xi})/P(\boldsymbol{\xi};\boldsymbol{\theta}) = P_{\mathbf{x}}(\boldsymbol{\xi}_{-i})/P(\boldsymbol{\xi}_{-i};\boldsymbol{\theta}) = \ldots = P_{\mathbf{x}}(\boldsymbol{\xi}^0)/P(\boldsymbol{\xi}^0;\boldsymbol{\theta}) = c, \qquad \text{(B.5)}
$$

where c is a constant that does not depend on $\boldsymbol{\xi}$ (actually, it does not depend

on $\boldsymbol{\xi}^0$ either). Thus, we have

$$P_{\mathbf{x}}(\boldsymbol{\xi}) = cP(\boldsymbol{\xi};\boldsymbol{\theta}) \text{ for all } \boldsymbol{\xi}. \tag{B.6}$$

On the other hand, both $P_{\mathbf{x}}$ and $P(.;\boldsymbol{\theta})$ are properly normalized probability distributions. Thus, we must have $c = 1$ because otherwise their sums over $\boldsymbol{\xi}$ could not both equal 1. This proves that if $J_{RM} = 0$, then $P(\boldsymbol{\xi};\boldsymbol{\theta}) = P_{\mathbf{x}}(\boldsymbol{\xi})$ for all $\boldsymbol{\xi}$. Using the identifiability assumption, this implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Thus, we have proven that $J_{RM}(\boldsymbol{\theta}) = 0$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The converse is trivial.

To prove consistency, we apply the law of large numbers. As sample size approaches infinity, $\tilde{J}_{RM}$ converges to J_{RM} (in probability, and up to the irrelevent additive constant). Thus, the estimator converges to a point where J_{RM} is globally minimized. By the proof just given, the global minimum is unique and found at the true parameter value (obviously, J_{RM} cannot be negative).

C Derivation of (20)

Noting that $g'(u) = -g^2(u)$, we can compute the gradient with respect to one element of $\mathbf{M}$ as

$$\begin{aligned}\nabla_{m_{ij}}\tilde{J} &= \hat{E}\sum_{i=1}^{n} -2(-2x_ix_j)\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i) \\ &\quad \times g^2(\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i)) \times g(\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i)) \\ &= \hat{E}x_ix_j\frac{4\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i)}{[1+\exp(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i)]^3} = 4\hat{E}x_ix_jh(-2x_i\mathbf{m}_i^T\mathbf{x} - 2b_ix_i),\end{aligned} \tag{C.1}$$

with

$$h(u) = \frac{\exp(u)}{(1+\exp(u))^3}. \tag{C.2}$$

We can further manipulate h:

$$\begin{aligned}h(u) &= \frac{\exp(-u/2)}{(\exp(u/2)+\exp(-u/2))^3} \\ &= \frac{1}{16\cosh^2(u/2)}\frac{\exp(u/2)+\exp(-u/2)-(\exp(u/2)-\exp(-u/2))}{\cosh(u/2)} \\ &= \frac{1}{8\cosh^2(u/2)}[1-\tanh(u/2)] = \frac{1}{8}(1-\tanh^2(u/2))(1-\tanh(u/2)).\end{aligned} \tag{C.3}$$

Now, let us note that $x_i\tanh(x_iu) = \tanh(u)$ for any $x = \pm 1$ because tanh is an odd function. Thus, the multiplying x_i inside the tanh function disappears together with the pre-multiplying x_i. Also, $\tanh^2(x_iu) = \tanh^2(u)$. Using these we finally obtain (20).

D Proof of Theorem 3

The proof is a simple variant of the partial integration trick used in basic score matching [3] based on earlier work by [7,8]. Simple manipulations of J_{NN} is (26) give

$$J_{NN} = -\int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})(\boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi}) \otimes \boldsymbol{\xi})^T(\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) \otimes \boldsymbol{\xi})d\boldsymbol{\xi} + \frac{1}{2}\int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})\|\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) \otimes \boldsymbol{\xi}\|^2 d\boldsymbol{\xi} + \text{const.} \quad \text{(D.1)}$$

where the constant only depends on $p_{\mathbf{x}}$. The latter term is clearly equal to the last term in (27), so what really needs to be proven is that the former term in (D.1) equals the sum of the first two terms in (27). The dot-product in that term consists of sums of term of the form

$$\int_{\mathbb{R}^n_+} p(\boldsymbol{\xi})\psi_{\mathbf{x},i}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2 d\boldsymbol{\xi}, \quad \text{(D.2)}$$

where $\psi_{\mathbf{x},i}$ denotes the i-th element of $\boldsymbol{\psi}_{\mathbf{x}}$. Now, we use partial integration as follows:

$$\begin{aligned}
\int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})\psi_{\mathbf{x},i}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2 d\boldsymbol{\xi} &= \int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})\frac{\partial_i p_{\mathbf{x}}(\boldsymbol{\xi})}{p_{\mathbf{x}}(\boldsymbol{\xi})}\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2 d\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^n_+} \partial_i p_{\mathbf{x}}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2 d\boldsymbol{\xi} \\
= p_{\mathbf{x}}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2\big|_{\xi_i=\infty} &- p_{\mathbf{x}}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2\big|_{\xi_i=0} - \int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})\partial_i(\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2)d\boldsymbol{\xi} \\
&= -\int_{\mathbb{R}^n_+} p_{\mathbf{x}}(\boldsymbol{\xi})[2\xi_i\psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) + \xi_i^2\partial_i\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})]d\boldsymbol{\xi}, \quad \text{(D.3)}
\end{aligned}$$

where the disappearance of the two terms in the last equality is due to the regularity assumptions of the theorem. Thus we have shown the theorem. A more rigorous justification for this partial integration element by element is given in Lemma 1 of [3].

Note that if we tried to do the same for the original score matching function, the term $p_{\mathbf{x}}(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi};\boldsymbol{\theta})\xi_i^2|_{\xi_i=0}$ would not contain the factor ξ_i^2, and it would not be zero in general, so there would be an extra term which is not very easy to evaluate (it contains the data pdf).

E Proof of Theorem 4

The assumption $q(\boldsymbol{\xi};\boldsymbol{\theta}^*) > 0$ implies $p_{\mathbf{x}}(\boldsymbol{\xi}) > 0$ almost everywhere in $\mathbb{R}^n_+$. Assume $J_{NN}(\boldsymbol{\theta}) = 0$. This implies that $\boldsymbol{\psi}_{\mathbf{x}}(.)$ and $\boldsymbol{\psi}(.;\boldsymbol{\theta})$ are equal a.e. because their squared distance is zero with a weight that is > 0 a.e. Thus, $\log p_{\mathbf{x}}(.) = \log p(.;\boldsymbol{\theta}) + c$ for some constant c. But c is necessarily 0 because both $p_{\mathbf{x}}$ and $p(.;\boldsymbol{\theta})$ are pdf's. Thus, $p_{\mathbf{x}} = p(.;\boldsymbol{\theta})$. By assumption, only $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ fulfills this equality, so necessarily $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and we have proven the implication from left to right. The converse is trivial.

References

[1] Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. *J. of the Royal Statistical Society ser B*, 36(2):192–236.

[2] Hinton, G. E., 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

[3] Hyvärinen, A., 2005. Estimation of non-normalized statistical models using score matching. *J. of Machine Learning Research*, 6:695–709.

[4] Ackley, D. H., Hinton, G. E., Sejnowski, T. J., 1985. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

[5] Hyvärinen, A., in press. Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*.

[6] Dawid, A. P., Lauritzen, S. L., 2005. The geometry of decision theory. In *Proc. 2nd Int. Symposium on Information Geometry and its Applications*, Tokyo, Japan.

[7] Pham, D.-T., Garrat, P., 1997. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725.

[8] Taleb, A., Jutten, C., 1999 Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820.