Relating ultrasonic vocalizations from a pair of rats to individual behavior: A composite link model approach

Nadia J. Vendrig1 | Lia Hemerik1 | Ilona J. Pinter2 | Cajo J.F. ter Braak1

1Biometris, Wageningen University & Research, Wageningen, 6700 AA Gelderland, The Netherlands
2Delta Phenomics BV, Nistelrooise Baan 3 5374 RE, Schaijk, Noord-Brabant, The Netherlands

Correspondence
Cajo J.F. ter Braak, Biometris, Wageningen University & Research, P.O. Box 16, 6700 AA Wageningen, Gelderland, The Netherlands.
Email: Cajo.terbraak@wur.nl

Funding information
Dutch Fund for Economic Structure Reinforcement, Grant/Award Number: 0908 (the "NeuroBasic PharmaPhenomics project")

Ultrasound vocalizations (USVs) are crucial in the social behavior of rats. We aim to relate USV rates of pairs of rats to individual activity in an automated home cage (PhenoTyper®) where USVs are recorded per pair and not per individual. We propose a composite link model approach to parametrize a mechanistic “sum-of-rates” model in which the pair's USV rate is the sum of the USV rates of individuals depending on their own behavior. In generalized linear models (GLMs), the individual's USV rates are multiplied. We verified through simulation that composite link model gave lower Poisson deviance than GLM. We analyzed the data from an experiment in which half of the cages did allow the pairs to interact (Pair Housing) and the other half did not (Individual Housing). The “sum-of-rates” model fits best for Individual Housing and GLM for Pair Housing. An additional simulation study strongly suggests that interaction between rats changes the underlying mechanism for vocalization behavior.

KEYWORDS
automated home cage, composite link model, model identification, social behavior, ultrasonic vocalization

1 INTRODUCTION

Automated home-cage systems have made it possible to study spontaneous behavior of laboratory rodents. Studying laboratory rodents (here, rats) in pairs is an essential step toward understanding their social behavior and toward using rats as a model species for studying...
psychiatric disorders with social impairment such as autism and depression. Recently, important steps have been made to improve the individual tracking of multiple rats in a home cage. These advances allow researchers to study vital social behavior such as playing and fighting. The PhenoTyper® system (Noldus, Wageningen, The Netherlands) tracks the whereabouts of both rats in the cage using a top-view camera and obtains their exact location in each video frame. The distance that the rat covers in the interval between two frames (further: time interval, typically 1/25 s) is used to calculate several parameters such as its velocity. With such a system, we can study multidimensional aspects of behavior (e.g., repetitive and stereotypic behaviors and social interaction, which are relevant to autism) for longer periods of time without human interference (Kas et al., 2014).

In addition to the movement of rats, ultrasonic vocalizations (USVs) are assumed to be an important part of rat behavior. USVs are indicators of the emotional state of the vocalizing rat (Brudzynski, 2009; Burgdorf et al., 2008), important for establishing and maintaining social contact such as in sexual and playing behavior (Himmler, Kisko, Euston, Kolb, & Pellis, 2014; Wöhr & Schwarting, 2013), and USVs can invoke response of other rats, for example, transmission of fear (Kim, Kim, Covey, & Kim, 2010). Emission of USVs in social settings is one of the most widely used means to study social communication in rodents (Servadio, Vanderschuren, & Trezza, 2015) and can provide more insight into the functional meaning of social behavior (Peters, Pothuizen, & Spruijt, 2015).

Using the Sonotrack® recording system (Metris, Hoofddorp, The Netherlands), we can continuously record vocalizations of rats in the PhenoTyper®. Interpretation of these data streams for studying autism, for example, requires development of analytical methods to establish the relation between USVs and behavior (Kas et al., 2014). As of yet, it has not been possible to assign recorded calls to individual rats when multiple rats are housed in the PhenoTyper®.

Allocating USVs to individual rats is difficult because the “voices” of rats do not noticeably differ between individuals and because USVs recorded in (automated home) cages cannot be traced back to their location of origin because echoes of USVs can be as loud as or louder than the original USVs (R. Bulthuis, Metris, personal communication 11th of October, 2016).

Being able to record activity per individual rat and not being able to record USVs per individual rat poses a challenge when integrating both data streams. Because of this challenge, Ågmo and Snoeren (2015), for example, could not analyze the full data of their experiment on the effect of USVs on mating behavior. We address the challenge by proposing a simple mechanistic model: Sound is recorded when either one of the rats produces sound, and the recorded vocalization rate is thus the sum of the two individual rates (“sum-of-rates” model). For the individual rates in relation to the individual activity, we then assume a generalized linear model (GLM). In this way, the model becomes a composite link model (CLM) (Thompson & Baker, 1981), which we extend to be applicable to fine-scaled video frame data in which the USVs are binary.

In the first part of the paper, we show that the “sum-of-rates” model performs better in simulations than a traditional Poisson GLM if the mechanistic model is true. In the second part of the paper, the “sum-of-rates” model is applied to a case study. In the case study, data from pairs of rats that could or could not interact are analyzed. In the mechanistic model, rats in a pair are considered to be two separate entities without interaction. We show that the potential for interaction between rats fundamentally changes the relation between activity and vocalization behavior, and thus, the mechanistic model proposed in the first part of the paper no longer holds when rats have been given the potential to interact, which strengthens the claim that vocalization behavior plays a role in social interaction.
### TABLE 1  Input data for the six-parameter combined activity model

| Activity Combination       | USVs      | Frequency |
|---------------------------|-----------|-----------|
| Stopping–Stopping         | $y_{SS}$  | $F_{SS}$  |
| Stopping–Lingering        | $y_{SL} + y_{LS}$ | $F_{SL} + F_{LS}$ |
| Stopping–Progressing      | $y_{SP} + y_{PS}$ | $F_{SP} + F_{PS}$ |
| Lingering–Lingering       | $y_{LL}$  | $F_{LL}$  |
| Lingering–Progressing     | $y_{LP} + y_{PL}$ | $F_{LP} + F_{PL}$ |
| Progressing–Progressing   | $y_{PP}$  | $F_{PP}$  |

Note. USVs = ultrasonic vocalizations.

## 2  STATISTICAL AND MECHANISTIC MODELING

### 2.1  Problem definition

We wish to integrate USVs and the activity of two rats. For simplicity, let us suppose that the activity per rat ($A_1$ and $A_2$ for rat 1 and 2, respectively) is recorded for every time interval in three categories: $S$, $L$, and $P$, denoting Stopping (sitting still), Lingering (moving slowly), and Progressing (moving quickly), respectively. For every time interval, we evaluate whether or not a USV was detected. There exist nine combinations of activity states ($A_1A_2 \in \text{States}$, with $\text{States} = \{SS, SL, SP, LS, LL, LP, PS, PL, PP\}$). For all these combinations, we can sum the time intervals with observed USVs ($y_{A_1A_2}$ with $A_1A_2 \in \text{States}$) and the total number of time intervals ($F_{A_1A_2}$ with $A_1A_2 \in \text{States}$). We are interested in estimating the cage-USV rate ($\mu_{A_1A_2}$ with $A_1A_2 \in \text{States}$) (i.e., the expected number of USVs per time interval). The cage-USV rate can easily be estimated using a Poisson GLM with responses $y_{A_1A_2}$, a log-link function, nine parameters, and an offset equal to the logarithm of $F_{A_1A_2}$.

If the vocalization rates of rats are the same, then $\mu_{SL} = \mu_{LS}$, $\mu_{SP} = \mu_{PS}$, and $\mu_{LP} = \mu_{PL}$, so that the number of parameters reduces from nine to six. In that case, the data can be represented as in Table 1 and the corresponding model is as follows:

**Combined Activity:**

$$\mu_{A_1A_2} = h(\beta_{A_1A_2}) = \exp(\beta_{A_1A_2}),$$  \hspace{1cm} (1)

where $h$ is the inverse of the log-link function ($h(x) = \exp(x)$), and $\beta_{A_1A_2}$ represents the parameter to be estimated. This Poisson GLM approach (Combined Activity GLM) is easy to implement and flexible, but we have to estimate a separate parameter for every combination of activity states and we completely disregard the fact that there exist two individual rats that both vocalize. Note that, due to the log-link function, the Combined Activity model estimates $\mu_{A_1A_2}$ as a product of (what we would like to interpret as) rates, where the rates are the exponentiated parameters. However, neither parameter has an interpretation as a rate, because rates should be summed rather than multiplied. For example, this model gives no estimate for the USV of a lingering rat.

An alternative Poisson GLM model (Count GLM) is to use the activity states per rat as explanatory variables, rather than activity combinations. If the vocalization rates of both rats are identical, we require only three parameters ($\beta_S$, $\beta_L$, and $\beta_P$) and can represent the data as in Table 2.

**Count GLM:**

$$\mu_{A_1A_2} = h(\beta_A + \beta_{A_2}) = \exp(\beta_{A_1} + \beta_{A_2}) = \exp(\beta_{A_1}) \cdot \exp(\beta_{A_2}).$$  \hspace{1cm} (2)

The Count GLM model reduces the number of parameters from six to three. It still disregards the fact that there are two individual rats that both vocalize. Note that, due to the log-link function, the Count GLM model estimates $\mu_{A_1A_2}$ as a product of (what we would like to interpret as) rates, where the rates are the exponentiated parameters. However, neither parameter has an interpretation as a rate, because rates should be summed rather than multiplied. For example, this model gives no estimate for the USV of a lingering rat.
TABLE 2  Input data for the three parameter Count generalized
linear model model

| Stopping | Activity  | Progressing | USVs | USV    | Frequency |
|----------|-----------|-------------|------|--------|-----------|
| 2        | 0         | 0           | ySS  | FSS    | yLL FLL   |
| 0        | 2         | 0           | yPP  | FPP    |           |
| 1        | 1         | 0           | ySL + yLS | FSL + FLS |
| 1        | 0         | 1           | ySP + yPS | FSP + FPS |
| 0        | 1         | 1           | yLP + yPL | FLP + FPL |

Note. USVs = ultrasonic vocalizations.

2.2  CLM approach

Mechanistically, the USV rate of a pair of rats is the sum of the USV rates of both rats. Therefore, a more mechanistic model would estimate the USV rates per rat based on their activity and would sum the following estimates:

\[ \mu_{A_1A_2} = \exp(\beta_{A_1}) + \exp(\beta_{A_2}). \]  \hspace{1cm} (3)

This “sum-of-rates” model can be recognized as a CLM (Thompson & Baker, 1981). In CLM, one observation can be linked to multiple linear predictors, which allows us to link the number of time intervals with observed USVs of a cage to a separate linear predictor per rat. The exponents of the three estimated parameters, \( \exp(\beta_S) \), \( \exp(\beta_L) \), and \( \exp(\beta_P) \), can be interpreted directly as the USV rate of a rat in the Stopping, Lingering, or Progressing activity state, respectively. An alternative mechanistic interpretation for the parameters of the same statistical model is described in Appendix A.

The concept of the CLM is easiest to grasp in matrix notation. The CLM model is written as follows:

\[ \mu = C h(X\beta), \]

where \( C \) is a so-called link matrix, \( h \) is the inverse of a link function (log-link in our situation), matrix \( X \) contains the observations, and vector \( \beta \) contains the parameters. If the link matrix is the identity matrix, the model is equivalent to a GLM. The “sum-of-rates” model is a CLM where \( C \) consists of two adjacent identity matrices and \( X \) has double the number of observations, one for every rat rather than one for every cage (see Appendix B for details).

In the special case of a model where the linear predictors are the same for both rats, the “sum-of-rates” model is equivalent to a Poisson GLM with a constant difference of \( \log(2) \). For example, a CLM model for predicting USV rate as a function of cage temperature \( T \) is as follows:

\[ \mu_T = \exp(\beta_i + \beta_T T) + \exp(\beta_i + \beta_T T) = 2 \exp(\beta_i + \beta_T T) = \exp(\log(2) + \beta_i + \beta_T T) = \exp(\beta_i^* + \beta_T T), \]  \hspace{1cm} (4)

where \( \beta_i \) is the intercept in the “sum-of-rates” model, and \( \beta_i^* \) is the updated intercept in the Poisson GLM approach.

2.3  Extension to multiple cages

Pairs of rats may vocalize with different overall frequencies. When extending our approach for multiple cages, we would therefore like to allow for a cage-specific intercept for each of the \( K \) cages \((k = 1, \ldots, K)\). As a consequence, one of the activity states (Stopping) is chosen as the reference
category, and $\beta_S$ is set to zero. In the GLM approach, the exponent of the cage-specific intercept $\beta_k$ can be interpreted as the USV rate of the cage when both rats are in the activity state Stopping. For the CLM approach (further: Count CLM), $\beta_k$ is included in the linear predictor of both rats, and $\exp(\beta_k)$ can thus be interpreted as the USV rate of one rat in the activity state Stopping. This extension is used in the case study below.

2.4 From counts to binary data

We record USVs as a binary variable per time interval (video frame): $y_{A_1A_2}$ for a time interval is either 1 or 0, depending on whether USV is detected or not within the time interval. However, the approach we have taken so far is based on counts observed in intervals: It estimates a USV rate assuming that $y_{A_1A_2}$ follows a Poisson distribution with a USV rate $\mu_{A_1A_2}$ per time interval and an interval length $F_{A_1A_2}$. The problem with the count approach is that it theoretically assumes that multiple USVs could occur within one time interval. This problem can be overcome by modeling the binary response per time interval using a binomial distribution with a complementary log–log-link function, which can be derived naturally from a truncated Poisson distribution (McCullagh & Nelder, 1989). For ease of notation, we drop the indices $A_1A_2$ for the moment. In short, the probability that $y = 1$ in a time interval is equal to the probability of detecting at least one USV, that is, the probability that the number of vocalizations ($N$) is not equal to zero. As $N$ is Poisson distributed,

$$P(y = 1) = P(N > 0) = 1 - P(N = 0) = 1 - \exp(-\mu t),$$  

(5)

where $\mu$ is the vocalization rate, and $t$ is the duration of the time interval. In a GLM model for binary data (binary GLM model), this equation results in the complementary log–log-link function ($H(x) = \log(-\log(1 - x))$), rather than the customary logistic function (McCullagh & Nelder, 1989).

We note for our model that the vocalization rate in Equation 5 is the total rate, which is the sum of the rates of the two individual rats (Equation 3). The resulting model can be expressed as a CLM model with two link functions, an inverse log-link that relates predictors to the rates and an inverse-complementary log–log-link function (Equation 5) that links the total rate to the binary observation as follows:

Binary sum-of-rates CLM:

$$P(y = 1) = E(y) = h_1(C(h(X\beta))),$$  

(6)

where $h_1(x) = 1 - \exp(-x)$ and $h(x) = \exp(x)$. This “sum-of-rates” model for binary data, briefly referred to as the binary CLM model, is a special case of the bilinear CLM of Thompson and Baker (1981).

3 SIMULATION STUDY

3.1 Material and methods

We simulate an experiment with one cage of two rats according to the binary version of the “sum-of-rates” model as in Section 2.4.

For each rat, activity data were generated as a Markov process with the three activity states as states using an input transition matrix based on the case study. USVs were generated per rat, using a Poisson process with a length of one time interval, and thereafter combined, yielding a “0” if no
USVs were detected in the time interval and a “1” otherwise (the data are thus truncated at 1). The input USV rates for the Poisson process per activity state are 0.01, 0.14, and 0.04 for Stopping, Lingering, and Progressing, respectively. These USV rates were loosely based on the real data, namely, the Pair Housing group of the case study (Section 4). Every data set was analyzed using a Combined Activity GLM, a count and a binary GLM approach, and a count data and a binary CLM approach. The simulation study was repeated using 10 times higher input USV rates.

Performance of the five models was quantified in terms of deviation of the model predictions from the expected values based on the true underlying data distribution. More specifically, we calculated the Poisson deviance (PD) of the model predictions as follows:

$$PD = 2 \sum_{A_1 \in \{S,L,P\}} \sum_{A_2 \in \{S,L,P\}} \left( f_{A_1A_2} \cdot \log \left( \frac{f_{A_1A_2}}{g_{A_1A_2}} \right) - f_{A_1A_2} + g_{A_1A_2} \right),$$

where $f_{A_1A_2}$ is the expected number of time intervals with USVs based on the true underlying data distribution, and $g_{A_1A_2}$ is the model prediction of a number of USVs. We also calculated the PD of the observed data. Source code to reproduce the results is available as an online appendix (Simulation 1).

### 3.2 Results

The Poisson deviances of the CLM models were lower than those of the corresponding GLM models, whereas the Poisson deviance of the Combined Activity GLM was in between (Figure 1, top row). With 10 times higher input USV rates, the binary models performed much better than the count models (Figure 1, bottom row). This was expected, as with higher USV rates, multiple USVs per time interval become more frequent than with lower USV rates, and the binary approach takes account of the discrepancy between the number of USVs (as produced by the rats) and the number of time intervals with USVs (as in the data set), and the number of USVs.

The simulation results also confirm that the activity state coefficients of the CLM models can be directly interpreted as USV rate per rat. These estimates are close to the input USV rates. Over all replications, all activity states, and both scenarios, the maximum observed squared error of a USV rate estimate of the binomial CLM model was less than 0.001.

### 4 CASE STUDY

#### 4.1 Experimental design and analysis

We observed 16 pairs of rats in automated home cages (PhenoTyper®, Noldus, Wageningen, The Netherlands) equipped with a Sonotrack® recording system (Metris, Hoofddorp, The Netherlands). In the first eight home cages, the two rats were separated by a wall in the middle, which did not allow interaction (Individual Housing). In the second eight home cages, the two rats were housed together and could interact (Pair Housing). For each home cage, we recorded whether or not a USV was observed and also the activity of each of the two rats (Stopping, Lingering, or Progressing) for a period of 30 min at a resolution of 25 video frames per second.

The number of observed USVs differed between the cages and was higher for Pair Housing than for Individual Housing. The number of USVs ranged from 5 to 368 for Individual Housing and from 164 to 728 for Pair Housing (Table 3). The observed vocalization rates tended to be larger when the rats were more active, and were highest when both rats were progressing. The activity

...
FIGURE 1  Poisson deviance of observed and modeled number of ultrasonic vocalizations on simulated data with realistic vocalization rates (top row) and data with increased vocalization rates (bottom row). The Combined Activity (Comb. Act.) generalized linear model (GLM) model has six parameters, and the count and binary models have three parameters. CLM = composite link model

patterns within cages differed between Individual and Pair Housing. The rats in Pair Housing spend more time being equally active ($F_{SS}$, $F_{LL}$, $F_{PP}$) than the rats in Individual Housing (Figure 2).

Data were analyzed using the following four candidate binomial CLM models:

**Intercept-only** The USV rate per time interval differs between the cages and is independent of activity of the rats in that cage.

$$\mu_k = \exp(\beta_k) + \exp(\beta_k).$$  \hspace{1cm} (8)

where $\mu_i$ is the USV rate per time interval for each of the $K$ cages ($k = 1, \ldots, K$), and $\beta_k$ is the intercept term per cage.
TABLE 3  The number of observed ultrasonic vocalizations (USVs) and the observed USV rate per 100 time intervals (denoted by $m$ subscripted with behavioral categories as in Table 1)

| Total USVs | $m_{SS}$ | $m_{SL}$ | $m_{SP}$ | $m_{LS}$ | $m_{LL}$ | $m_{LP}$ | $m_{PS}$ | $m_{PL}$ | $m_{PP}$ |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Individual Housing | | | | | | | | | |
| Cage 1 | 40 | 0.045 | 0.087 | 0.048 | 0.016 | 0.093 | 0.086 | 0.150 | 0.261 | 0.000 |
| Cage 2 | 25 | 0.011 | 0.059 | 0.146 | 0.027 | 0.026 | 0.087 | 0.188 | 0.176 | 0.323 |
| Cage 3 | 5 | 0.014 | 0.000 | 0.000 | 0.014 | 0.014 | 0.000 | 0.000 | 0.038 | 0.000 |
| Cage 4 | 368 | 0.751 | 0.827 | 0.939 | 0.673 | 0.821 | 0.801 | 1.023 | 0.967 | 0.756 |
| Cage 5 | 327 | 0.847 | 0.812 | 0.854 | 0.422 | 0.574 | 0.334 | 1.374 | 2.163 | 1.708 |
| Cage 6 | 31 | 0.040 | 0.020 | 0.066 | 0.075 | 0.055 | 0.068 | 0.158 | 0.210 | 0.402 |
| Cage 7 | 391 | 0.471 | 0.692 | 0.997 | 0.479 | 0.815 | 0.942 | 1.091 | 1.224 | 1.675 |
| Cage 8 | 186 | 0.332 | 0.350 | 0.958 | 0.314 | 0.232 | 1.030 | 0.476 | 0.285 | 1.311 |
| Pair Housing | | | | | | | | | |
| Cage 1 | 164 | 0.106 | 0.233 | 0.875 | 0.198 | 0.317 | 0.911 | 0.402 | 0.946 | 1.125 |
| Cage 2 | 347 | 0.523 | 0.481 | 0.912 | 0.363 | 0.464 | 1.165 | 1.187 | 0.927 | 2.659 |
| Cage 3 | 728 | 0.939 | 1.204 | 1.557 | 0.855 | 1.102 | 1.708 | 1.982 | 2.528 | 4.616 |
| Cage 4 | 232 | 0.264 | 0.360 | 1.300 | 0.191 | 0.363 | 0.546 | 0.428 | 0.878 | 1.653 |
| Cage 5 | 290 | 0.116 | 0.353 | 2.025 | 0.128 | 0.249 | 1.177 | 1.048 | 1.603 | 2.301 |
| Cage 6 | 337 | 0.079 | 0.148 | 0.450 | 0.181 | 0.334 | 0.744 | 1.896 | 1.915 | 5.517 |
| Cage 7 | 583 | 0.404 | 0.509 | 1.244 | 0.461 | 0.989 | 1.974 | 3.239 | 2.867 | 5.513 |
| Cage 8 | 615 | 0.770 | 0.493 | 2.790 | 0.506 | 0.601 | 1.744 | 1.949 | 2.018 | 7.603 |

FIGURE 2  Observed number of time intervals for all combinations of Stopping, Lingering, and Progressing per cage. For each cage, the left bar indicates the time intervals in which both rats were equally active (from bottom to top: $F_{SS}$ (yellow), $F_{LL}$ (pink), and $F_{PP}$ (blue)), and the right bar indicates the time intervals in which the rats were not equally active (from bottom to top: $F_{SL}$ and $F_{LS}$ (orange); separated by black bar), $F_{SP}$ and $F_{PS}$ (green; separated by black bar), and $F_{LP}$ and $F_{PL}$ (purple; separated by black bar).

Current Activity  The USV rate per time interval differs between the cages and is dependent on the activity of rat 1 and activity of rat 2 in that frame.

$$\mu_{kA_1A_2} = \exp(\beta_k + \beta_{L_{X_1}} + \beta_{P_{X_2}}) + \exp(\beta_k + \beta_{L_{X_2}} + \beta_{P_{X_1}}),$$  (9)
where $\mu_{kA_1A_2}$ is the USV rate per time interval for each of the $K$ cages ($k = 1, \ldots, K$) given the activity of rat 1 ($A_1 = S, L,$ or $P$) and rat 2 ($A_2 = S, L,$ or $P$); $\beta_k$ is the intercept term per cage; $\beta_L$ and $\beta_P$ are the regression coefficients for Lingering and Progressing, respectively; and $X_{L_1}, X_{L_2}, X_{P_1},$ and $X_{P_2}$ are indicator variables that indicate whether or not rat 1 ($X_{A_1}$) and rat 2 ($X_{A_2}$) are lingering ($X_{L_n}$) or progressing ($X_{P_n}$). Note that, as is customary for regression models with nominal variables (factors), the intercept per cage here is the expected USV rate per time interval when the rat is Stopping (i.e., the baseline USV rate), and the regression coefficients for Lingering measure the difference in log-vocalization rate between Lingering and Stopping and similarly for Progressing.

**Past Activity** The USV rate per time interval differs between the cages and is dependent on the activity of rat 1 and activity of rat 2 in that time interval and the 24 time intervals before (1 s). The model equation is the same as Equation 9 with a different definition of all parameters except $\beta_k, \beta_L,$ and $\beta_P.$ In the Past Activity model, $\mu_{ij}$ is the USV rate per time interval for each of the $K$ cages ($i = 1, \ldots, K$) and $F$-24 time intervals ($f = 25, \ldots, F$) given the activity of rat 1 and rat 2 in the second before that time interval, and $X_{L_1}, X_{L_2}, X_{P_1},$ and $X_{P_2}$ are the proportion of time interval in the previous second rat 1 ($X_{A_1}$) and rat 2 ($X_{A_2}$) were lingering ($X_{L_n}$) and progressing ($X_{P_n}$). The first 24 time intervals in the data set cannot be used because we do not have recorded activity in the second before.

**Averaged Past Activity** The USV rate per time interval differs between the cages and is dependent on the combined activity of both rats in the second (25 frames) before that time interval. The model equation is the same as Equation 9 with a different interpretation of all parameters except $\beta_k, \beta_L,$ and $\beta_P.$ In the Averaged Past Activity model, $\mu_{ij}$ is the USV rate per time interval for each of the $K$ cages ($i = 1, \ldots, K$) and $F$-24 time intervals ($f = 25, \ldots, F$) given the averaged activity of both rats in the second (25 video frames) before that time interval; $X_{L_1}$ equals $X_{L_2},$ and $X_{P_1}$ equals $X_{P_2}$ which are the proportion of time interval in the previous second in which rat 1 and rat 2 were lingering and progressing. The model equation can be rewritten as a GLM model where $\beta_k^* = \beta_k + \log(2)$ (as in Equation 4).

The best candidate model for the data set is chosen based on Akaike information criterion (AIC; the lower, the better). The Intercept-Only model had the highest AIC for both the Individual Housing and the Pair Housing data set, which confirms the hypothesized relation between activity and vocalization behavior (Table 4). The Individual Housing and Pair Housing data set have a different best candidate model. In the Individual Housing data, the best candidate model was the Current Activity model. This model is similar to the one used in the simulation study and is

| Model                    | Parameters | Individual Housing | Pair Housing |
|--------------------------|------------|--------------------|--------------|
|                           | AIC        | $\Delta$AIC | rank | AIC        | $\Delta$AIC | rank |
| Intercept-Only           | 8          | 15,852           | 4 | 30,819 | -102   | 4   |
| Current Activity         | 10         | **15,750**       | **1** | 29,104 | -1,715 | 3   |
| Past Activity            | 10         | 15,759           | -94 | 28,830 | -1,989 | 2   |
| Averaged Past Activity   | 10         | 15,756           | -96 | **28,700** | **-2,119** | **1** |

*Note.* Data of the model with the lowest AIC per data set are printed in bold. The residual degrees of freedom is comparable between models due to the very large number of time intervals per cage (roughly 45,000) compared with their number of parameters.
In this paper, we presented two approaches to relate the USV rate per cage to the activities of two rats, which we evaluated in a small simulation study. In the CLM approach, like in the mechanistic model based on the AIC differences between the Current Activity model and the (Averaged) Past Activity model were between 6 and 8, which is indicative of a trend. For the Pair Housing data, in which interaction between rats was possible, the Current Activity model was outperformed by the Averaged Past Activity Model ($\Delta AIC = 404$). The addition of the potential to interact thus resulted in vocalization behavior for which the mechanistic model as proposed in Section 2.4 is no longer valid. In the presence of interaction, the USVs are better predicted from the joined activity of the pair than from the activity of the individuals. Because the Averaged Past Activity Model has equal linear predictors for both rats in the pair, it has an equivalent GLM model (see Equation 4).

In addition to the abstract quantification of model quality obtained from AICs, we provide a more tangible measure for models comparison. We compare the average predicted probabilities to vocalize for time intervals with and without an observed USV for each cage (Figure 3) and interpret the difference between these probabilities as indicative for the quality of model predictions (higher difference, better predictions). The estimates of the Intercept-Only model per cage are equal to the proportion of time intervals with an observed USV. For the other models, the differences in predicted vocalization probabilities within models were larger for Pair Housing than for Individual Housing (relative to the differences between cages). The differences between models within pairs were also larger for Pair Housing than for Individual Housing; the Averaged Past Activity model was clearly superior in the Pair Housing data, whereas the Current Activity model was not clearly superior in the Individual Housing data.

The data and source code to reproduce the results are available as online supporting information (within Simulation 2).

5 | DISCUSSION

In this paper, we presented two approaches to relate the USV rate per cage to the activities of two rats, which we evaluated in a small simulation study. In the CLM approach, like in the mechanistic
model, the activity of a rat predicts its USV rate, and the sum of USV rates of both rats gives the USV rate per cage ("sum-of-rates"). In contrast, in the GLM approach, the combined activity of both rats predicts the USV rate per cage. On data simulated using the mechanistic model, the Count and Binary CLM models were better able to predict the USV rate per cage than the Count and Binary GLM. In our design, we recorded which time intervals contained a USV rather than recording the USVs itself, which explains why the Binary models predicted the USV rates better than their Count counterparts. Overdispersion was not an issue, as the USV counts were presumably near binary, and binary data cannot have overdispersion.

Application of the CLM approach in a case study on the data of rats that could not interact (the Individual Housing data) resulted in the "sum-of-rates" model (Current Activity) as the best model in terms of AIC. When analyzing data from rats that could interact (Pair Housing data), however, we found that the "sum-of-rates" model did not provide the best model in terms of AIC. Instead, a CLM model with the averaged activity of both rats as predictor (Averaged Past Activity, which has an equivalent GLM model) fitted better.

The AIC is not a formal test but rather a guideline for model selection. Models with an AIC difference lower than 2 can be considered equivalent, whereas a difference over 10 implies that they are “definitely different” (Bolker, 2007). The AIC differences for Pair Housing were all over 100 and thus different. The AIC differences for Individual Housing among the Current Activity, Past Activity, and Averaged Past Activity model were between 2 and 9, which is merely indicative of a trend. In terms of model predictions (Figure 3), the differences between models were also apparent for Pair Housing and nonpersuasive for Individual Housing. For the Pair Housing data, the AIC differences rule out the Current Activity model. In contrast, for the Individual Housing data, the Current Activity model is the best model in terms of AIC, and the Averaged Past Activity model is a close runner-up. From a biological perspective, however, because the rats cannot interact, it seems highly unlikely that rats vocalize based on the behavior of the other. We verified that we have enough power in our analysis to select the “true” model via a second simulation study. In this simulation study, we alternately assume that one of the four candidate models is true and generate data accordingly. All generated data were analyzed using all four candidate models. In almost all instances, the candidate model with the lowest AIC in the simulation was the candidate model used to generate the data and vice versa (more details and results are in Appendix C). From this simulation study, we can thus conclude that it is highly unlikely that data with an underlying model structure from one of the candidate models result in another best candidate model.

We posed a mechanistic model for the vocalization rate observed in a cage in relation to the behavior of the individual rats. This model led to a CLM (Equation 3). We thus went from a mechanistic model to a statistical model. However, the mechanistic model is not the only one leading to this specific statistical model. Another mechanistic model would be that rat 1 vocalizes only in response to the behavior of rat 2 and vice versa. This leads also to the statistical model of Equation 3. For the full set of models, see Appendix A.

Many more models than those described here could be fitted to the data. Given that the USV rate varied widely between pairs, a mixed modeling approach could be an interesting future extension of our work. For mixed modeling approaches for CLM, see Ayma, Durbán, Lee, and Eilers (2016) and Douma (2017).

In spite of our limited selection of models, we can still infer that rats that can and rats that cannot interact show a different relation between USVs and activity. More specifically, the USV rates estimated from individual behavior of rats gives the best predictor when rats are housed individually, but when rats are housed in pairs, such an individualistic model no longer gives the best predictor, and a predictor based on the combination of the rats' behavior performs better.
The USV rate of a pair of interacting rats is thus shown to be different from the sum of its parts. We conclude from this experiment that social interaction between rats changes the relation between activity and USV rate of the rats.

With this application, we have once more demonstrated the utility of the CLM approach. A small R library implementing this approach (PBCLM-package) and the R code to replicate both simulation studies are available as online supporting information.

ACKNOWLEDGEMENTS
The research leading to these results has received funding from the Dutch Fund for Economic Structure Reinforcement (FES), under Grant 0908 (the “NeuroBasic PharmaPhenomics project”). We thank María Xosé Rodríguez-Álvarez and two anonymous reviewers for constructive comments.

ORCID
Nadia J. Vendrig http://orcid.org/0000-0002-3606-9769
Lia Hemerik http://orcid.org/0000-0002-6892-2840
Cajo J.F. ter Braak http://orcid.org/0000-0002-0414-8745

REFERENCES
Ágmo, A., & Snoeren, E. M. S. (2015). Silent or vocalizing rats copulate in a similar manner. *PloS One*, 10(12), 1–13.
Ayma, D., Durbán, M., Lee, D. J., & Eilers, P. H. (2016). Penalized composite link models for aggregated spatial count data: A mixed model approach. *Spatial Statistics*, 17, 179–198. http://doi.org/10.1016/j.spasta.2016.06.007
Bolker, B. M. (2007). *Ecological models and data in R*. Princeton, NJ: Princeton University Press.
Brudzynski, S. M (2009). Communication of adult rats by ultrasonic vocalization: Biological, sociobiological, and neuroscience approaches. *ILAR Journal*, 50(1), 43–50. http://ilarjournal.oxfordjournals.org/content/50/1/43.full.pdf
Burgdorf, J., Kroes, R. A., Moskal, J. R., Pfaus, J. G., Brudzynski, S. M., & Panksepp, J. (2008). Ultrasonic vocalizations of rats (*Rattus norvegicus*) during mating, play, and aggression: Behavioral concomitants, relationship to reward, and self-administration of playback. *Journal of Comparative Psychology*, 122(4), 357–367. http://www.ncbi.nlm.nih.gov/pubmed/19014259
Douma, D. I. (2017). *The penalized composite link mixed model: The estimation of the underlying de-aggregated distribution of an aggregated variable* (master thesis). Leiden University, The Netherlands.
Himmler, B., Kisko, T., Euston, D., Kolb, B., & Pellis, S. (2014). Are 50-khz calls used as play signals in the playful interactions of rats? I. Evidence from the timing and context of their use. *Behavioural Processes*, 106, 60–66. http://www.ncbi.nlm.nih.gov/pubmed/24811452
Kas, M. J., Glennon, J. C., Buitema, J., Ey, E., Biemans, B., Crawley, J., … Steckler, T. (2014). Assessing behavioural and cognitive domains of autism spectrum disorders in rodents: Current status and future perspectives. *Psychopharmacology*, 231(6), 1125–1146.
Kim, E. J., Kim, E. S., Covey, E, & Kim, J. J. (2010). Social transmission of fear in rats: The role of 22-khz ultrasonic distress vocalization. *PloS One*, 5(12), e15077.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: CRC Press.
Peters, S. M., Pothuizen, H. H. J., & Spruijt, B. M. (2015). Ethological concepts enhance the translational value of animal models. *European Journal of Pharmacology*, 759, 42–50. https://doi.org/10.1016/j.ejphar.2015.03.043
Servadio, M., Vanderschuren, L. J. M. J., & Trezza, V. (2015). Modeling autism-relevant behavioral phenotypes in rats and mice: Do 'autistic' rodents exist. *Behavioural Pharmacology*, 26, 522–40. http://www.ncbi.nlm.nih.gov/pubmed/26226143
SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Vendrig NJ, Hemerik L, Pinter IJ, ter Braak CJF. Relating ultrasonic vocalizations from a pair of rats to individual behavior: A composite link model approach. *Statistica Neerlandica*. 2019;73:139–156. [https://doi.org/10.1111/stan.12144](https://doi.org/10.1111/stan.12144)

APPENDIX A

ALTERNATIVE MECHANISTIC INTERPRETATION OF THE STATISTICAL MODEL

We posed a mechanistic model for the USV rate observed in a cage in relation to the behavior of the individual rats. This model led to a statistical model: a composite link model (Equation 3 in the main text). A valid question is whether the mechanistic model is the only one leading to this statistical model. In this section, we show that it is not.

In this paper, we have assumed a mechanistic model in which the two rats in the same cage have a USV rate depending on their activity state, which sums up to the USV rate of the cage. In this model, we can interpret $\exp(\beta_{A1})$ as the USV rate of the first rat and $\exp(\beta_{A2})$ as the USV rate of the second rat. This is equivalent to saying that $\exp(\beta_{A1})$ is the USV rate of the cage conditional on the activity of the first rat and that $\exp(\beta_{A2})$ is the USV rate of the cage conditional on the activity of the second rat.

However, this is not the only possible interpretation for the parameters from the statistical model. Now, let us assume a mechanistic model in which the USVs of rats are related not only to their own activity but also to the activity of the other rat in the cage. In the most extreme case, rat 1 would vocalize solely conditional on the activity of rat 2 and vice versa. Because we have no way of detecting which rat uttered which USV, this assumption would lead to exactly the same statistical model and exactly the same estimated model parameters. The full collection of models can be written as follows:

$$
\mu_{A1,A2} = \exp(p\beta_{A1} + (1-p)\beta_{A1}) + \exp(q\beta_{A2} + (1-q)\beta_{A2}) = \exp(\beta_{A1}) + \exp(\beta_{A2}),
$$

where $p$ and $q$ range from 1 when USVs are solely related to a rat's own behavior to 0 when USVs of a rat are solely related to the other rat's behavior (Table A1). The parameters $p$ and $q$ are inestimable when the observed USVs cannot be allocated to one of the rats.

Now, the USV rates of the cage, conditional on the activity of the first or second rat can still be estimated by $\exp(\beta_{A1})$ and $\exp(\beta_{A2})$, respectively. However, the USV rates of rat 1 and rat 2 are given by $\exp(p\beta_{A1} + (1-q)\beta_{A1})$ and $\exp((1-p)\beta_{A2} + q\beta_{A2})$, respectively, which we cannot estimate.
The conclusions that we draw in the main text are based on the contrast of models that differ in their relation between activity and vocalization for rats that can and cannot interact. This conclusion is not dependent on the precise mechanistic interpretation of the CLM model.

APPENDIX B

CLM MODEL

The concept of the CLM (Thompson & Baker, 1981) is easiest to grasp in matrix notation. In this Appendix, we first write the Count GLM model in CLM notation and thereafter the “sum-of-rates” model. For brevity, we will show matrices as if there exist only two activity states (S and L). In that case, the model matrices for the Count GLM model are as follows:

$$\mu = \begin{bmatrix} \mu_{SS} \\ \mu_{SL} \\ \mu_{LS} \\ \mu_{LL} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad X = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_S \\ \beta_L \end{bmatrix}.$$

Here, $C$ is the identity matrix and thus $Ch(X\beta) = h(X\beta)$,

so that $\mu = h(X\beta)$

$$= \exp\left( \begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} \beta_S \\ \beta_L \end{bmatrix} \right) = \exp\left( \begin{bmatrix} 2 \cdot \beta_S \\ \beta_S + \beta_L \\ \beta_L + \beta_S \\ 2 \cdot \beta_L \end{bmatrix} \right).$$

which is equivalent to Equation 2 (in the main text).

The link matrix allows us to link multiple observations to one linear predictor. In order to obtain the “sum-of-rates” model of Equation 3, we set matrices $X$ and $C$ to the following:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$
so that \( h(X\beta) = \exp \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_s \\ \beta_L \end{bmatrix} = \begin{bmatrix} \exp(\beta_S) \\ \exp(\beta_S) \\ \exp(\beta_L) \\ \exp(\beta_L) \\ \exp(\beta_L) \\ \exp(\beta_L) \\ \exp(\beta_S) \\ \exp(\beta_L) \end{bmatrix} \)

and

\[
\begin{bmatrix}
\mu_{SS} \\
\mu_{SL} \\
\mu_{LS} \\
\mu_{LL}
\end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \exp(\beta_S) \\ \exp(\beta_S) \\ \exp(\beta_S) \\ \exp(\beta_S) \end{bmatrix} + \begin{bmatrix} \exp(\beta_S) \\ \exp(\beta_S) \\ \exp(\beta_S) \\ \exp(\beta_S) \end{bmatrix}
\]

which is equivalent to Equation 3 (in the main text). For clarity, the separation between the observations is visualized using white space in the matrices.

**APPENDIX C**

**SIMULATION STUDY FOR AICs**

In the case study, we observed that out of four candidate models, the Current Activity model had the lowest AIC on the Individual Housing data, and the Average Past Activity model had the lowest AIC on the Pair Housing data. In this simulation study, we show that it is highly unlikely that, assuming that the true model is among the candidate models, the same model holds for the Individual Housing and Pair Housing data, and we show that AIC is a suitable method for model selection in this type of data. Source code to reproduce the results is available as online supporting information (Simulation 2).

**TABLE C1** For each of the eight simulation scenarios (two experimental data sets \( \times \) four candidate models) and for each of the four candidate models used for analysis, the proportion of 1,000 simulations in which each candidate model has the lowest Akaike information criterion

| Simulation data set   | Candidate model used for analysis | IC-Only | Cur. Act. | Past Act. | Ave. Past Act. |
|-----------------------|----------------------------------|---------|-----------|-----------|---------------|
| Individual Housing    | IC-Only                          | 0.805   | 0.098     | 0.037     | 0.060         |
|                       | Cur. Act.                        | 0.000   | 1.000     | 0.000     | 0.000         |
|                       | Past Act.                        | 0.000   | 0.000     | 0.708     | 0.292         |
|                       | Ave. Past Act.                   | 0.000   | 0.000     | 0.267     | 0.733         |
| Pair Housing          | IC-Only                          | 0.761   | 0.113     | 0.056     | 0.070         |
|                       | Cur. Act.                        | 0.000   | 1.000     | 0.000     | 0.000         |
|                       | Past Act.                        | 0.000   | 0.000     | 1.000     | 0.000         |
|                       | Ave. Past Act.                   | 0.000   | 0.000     | 0.000     | 1.000         |

*Note.* IC-Only = Intercept-Only; Cur. Act. = Current Activity; Past Act. = Past Activity; Ave. Past Act. = Average Past Activity.
In our simulation study, for every combination of the four candidate models and the two experimental data sets (Individual Housing and Pair Housing), we

(i) fit the candidate model to the experimental data set to obtain an Input Model,
(ii) predict the expected USV rate for every frame in the data set using the model predictions from the Input Model,
(iii) generate 1,000 simulation data sets using the expected USV rates from the previous step,
(iv) fit all four candidate models to the 1,000 generated simulation data sets, and
(v) compare the AICs of the four candidate models used for analysis and determine which candidate model has the lowest AIC, for every of the simulated data sets.

Note that a candidate model here indicates the structure of the statistical model, not the size of coefficients. When comparing AICs of the candidate models, we always use the four AICs calculated in step (iv) of the procedure. The correct candidate model is thus the candidate model that has the same structure as the Input Model, not the exact same model.

The results of the simulation study show that AIC is a reliable tool for selecting the correct candidate model. In the vast majority of 8,000 simulation data sets, the correct candidate model had the lowest AIC. This is more often true in the data based on the Pair Housing (95.6%) than in the Individual Housing data (81.2%), and it depends on which was the correct candidate model (Table C1). In the simulation data based on the Pair Housing case study, the correct candidate model always had the lowest AIC except when the correct candidate model was the Intercept-Only model. In the simulation data based on the Individual Housing case study, the Current Activity model always had the lowest AIC when it was correct.

Differences in AIC between the correct candidate model and the other three models (Figure C1) were larger in the Pair Housing than the Individual Housing set and differed between the different candidate models used to estimate the Input model. The order of magnitude of the AIC differences when the correct candidate model had the lowest AIC was the same as the AIC differences observed in the case study. The maximum AIC difference observed when the wrong candidate model was identified was 13.3, which is well below the observed AIC differences for the case study.

Secondly, we use the simulation to show that it is unlikely that the two data sets from the case study have the same underlying model, given that a different candidate model had the lowest AIC when analyzing this data (Current Activity for Individual Housing vs. Average Past Activity for Pair Housing). For this aim, we determine how likely it is that the candidate model with the lowest AIC is in fact the correct candidate model. If it is very likely that the candidate model with the lowest AIC is the correct candidate model, and the candidate models with the lowest AIC are not the same for both data sets, it is unlikely that the underlying “true” model is the same.

In the simulation study, in the Individual Housing data sets, when the AIC of the Current Activity model was the lowest, that candidate model was the true candidate model in 91% of simulation data sets (Table C2). The Current Activity model never had the lowest AIC when the Average Past Activity model was the true candidate model. In the Pair Housing simulation data sets, when the AIC of the Average Previous Second model is the lowest, that candidate model was the true candidate model in 96% of simulation data sets. The Average Past Activity model never had the lowest AIC when the Current Activity model was the true candidate model. It is thus both unlikely that in the case study, the true candidate model of the Individual Housing data set was the Average Past Activity model and that in the Pair Housing data set, the true candidate model was the Current Activity model.
FIGURE C1  Results of the simulation study for Akaike information criterion (AIC). Each panel represents results on 1,000 simulation data sets based on the Individual Housing (left column) or Pair Housing (right column) experiment, using each of the four candidate models (rows) to generate input data (i.e., rows represent the true models). Violin plots indicate the distribution of the difference in AIC between the Intercept-Only model (purple reference line) and the candidate model.
TABLE C2  For data simulated based on Individual Housing and Pair Housing, given that a candidate model has the lowest AIC, what is the observed probability that each of the four candidate models were used to generate the data set

| Simulation data set | Candidate model with lowest AIC |
|---------------------|---------------------------------|
|                     | IC Only | Cur. Act. | Past Act. | Ave. Past Act. |
| Individual Housing  |         |          |          |                |
| IC Only             | 1.000   | 0.089    | 0.037    | 0.055          |
| Cur. Act.           | 0.000   | 0.911    | 0.000    | 0.000          |
| Past Act.           | 0.000   | 0.000    | 0.700    | 0.269          |
| Ave. Past Act.      | 0.000   | 0.000    | 0.264    | 0.676          |
| Pair Housing        |         |          |          |                |
| IC Only             | 1.000   | 0.102    | 0.053    | 0.065          |
| Cur. Act.           | 0.000   | 0.898    | 0.000    | 0.000          |
| Past Act.           | 0.000   | 0.000    | 0.947    | 0.000          |
| Ave. Past Act.      | 0.000   | 0.000    | 0.000    | 0.935          |